

# Von Neumann Entropy Penalization and Low Rank Matrix Estimation

Vladimir Koltchinskii \*

School of Mathematics  
Georgia Institute of Technology  
Atlanta, GA 30332-0160  
vlad@math.gatech.edu

September 14, 2010

## Abstract

A problem of estimation of a Hermitian nonnegatively definite matrix  $\rho$  of unit trace (for instance, a density matrix of a quantum system) based on  $n$  independent measurements

$$Y_j = \text{tr}(\rho X_j) + \xi_j, \quad j = 1, \dots, n$$

is studied,  $\{X_j\}$  being i.i.d. Hermitian matrices and  $\{\xi_j\}$  being i.i.d. mean zero random variables independent of  $\{X_j\}$ .

The estimator

$$\hat{\rho}^\varepsilon := \operatorname{argmin}_{S \in \mathcal{S}} \left[ n^{-1} \sum_{j=1}^n (Y_j - \text{tr}(S X_j))^2 + \varepsilon \text{tr}(S \log S) \right]$$

is considered, where  $\mathcal{S}$  is the set of all nonnegatively definite Hermitian  $m \times m$  matrices of trace 1. The goal is to derive oracle inequalities showing how the estimation error depends on the accuracy of approximation of the unknown state  $\rho$  by low-rank matrices.

**Keywords and phrases:** low rank matrix estimation, von Neumann entropy, matrix regression, empirical processes, noncommutative Bernstein inequality, quantum state tomography

**2010 AMS Subject Classification:** 62J99, 62H12, 60B20, 60G15, 81Q99

---

\*Partially supported by NSF Grants DMS-0906880 and CCF-0808863

# 1 Introduction

Let  $\mathbb{M}_m(\mathbb{C})$  be the set of all  $m \times m$  matrices with complex entries and let

$$\mathcal{S} := \left\{ S \in \mathbb{M}_m(\mathbb{C}) : S = S^*, S \geq 0, \text{tr}(S) = 1 \right\}$$

be the set of all nonnegatively definite Hermitian matrices of trace 1. Here and in what follows  $S^*$  denotes the adjoint matrix of  $S$  and  $\text{tr}(S)$  denotes its trace. The matrices from the set  $\mathcal{S}$  can be interpreted, for instance, as *density matrices*, describing the states of a quantum system. Given a Hermitian matrix  $X$  (*an observable*), its expectation in a state  $\rho \in \mathcal{S}$  is defined as  $\mathbb{E}_\rho X := \text{tr}(\rho X)$ . Let  $X_1, \dots, X_n \in \mathbb{M}_m(\mathbb{C})$ ,  $X_j = X_j^*$ ,  $j = 1, \dots, n$  be given Hermitian matrices (observables) and let  $\rho \in \mathcal{S}$  be an unknown state of the system. An important problem in *quantum state tomography* is to estimate  $\rho$  based on the observations  $(X_j, Y_j)$ ,  $j = 1, \dots, n$ , where

$$Y_j = \text{tr}(\rho X_j) + \xi_j, \quad j = 1, \dots, n,$$

$\xi_j$ ,  $j = 1, \dots, n$  being i.i.d. random variables with mean zero and finite variance representing measurement errors. In other words, the unknown state  $\rho$  of the system is to be learned based on a set of measurements in a number of “directions”  $X_j$ ,  $j = 1, \dots, n$  (see Artiles, Gill and Guta (2004) for a general discussion of statistical problems in quantum state tomography). In what follows, it will be usually assumed that the design variables  $X, X_1, \dots, X_n$  are also random, specifically, they are i.i.d. Hermitian  $m \times m$  matrices with distribution  $\Pi$ , and they are independent of the noise  $\{\xi_j\}$ .

A typical choice of the design variables already discussed in the literature (see Gross et al (2009), Gross (2009)) can be described as follows. The linear space of matrices  $\mathbb{M}_m(\mathbb{C})$  can be equipped with the Hilbert-Schmidt inner product:  $\langle A, B \rangle := \text{tr}(AB^*)$ . Let  $E_i$ ,  $i = 1, \dots, m^2$  be an orthonormal basis of  $\mathbb{M}_m(\mathbb{C})$  consisting of Hermitian matrices  $E_i$ . Let  $X_j$ ,  $j = 1, \dots, n$  be i.i.d. random variables sampled from a distribution  $\Pi$  on the set  $\{E_1, \dots, E_{m^2}\}$ . We will refer to this model as *sampling from an orthonormal basis*. Most often, the uniform distribution  $\Pi$  that assigns probability  $m^{-2}$  to each basis matrix  $E_i$  will be used. Note that in this case  $\mathbb{E}|\langle A, X \rangle|^2 = m^{-2}\|A\|_2^2$ , where  $\|\cdot\|_2 := \langle \cdot, \cdot \rangle^{1/2}$  is the Hilbert-Schmidt (or the Frobenius) norm.

The following simple example is related to the problems of *matrix completion* extensively discussed in the recent literature (see, e.g., Candes and Recht (2009), Candes and Tao (2009), Recht (2009) and references therein). More precisely, it deals with a version

of matrix completion for Hermitian matrices (see Gross (2009)). In this case, when one knows an entry  $\rho_{ij}$  of a matrix  $\rho$ , one also knows the entry  $\rho_{ji} = \bar{\rho}_{ij}$ .

**Example 1. Matrix completion.** Let  $\{e_i : i = 1, \dots, m\}$  be the canonical basis of  $\mathbb{C}^m$ . Then, the following set of Hermitian matrices forms an orthonormal basis of  $\mathbb{M}_m(\mathbb{C})$  :

$$\left\{ e_i \otimes e_i : i = 1, \dots, m \right\} \cup \left\{ \frac{1}{\sqrt{2}}(e_i \otimes e_j + e_j \otimes e_i) : 1 \leq i < j \leq m \right\} \\ \cup \left\{ \frac{i}{\sqrt{2}}(e_i \otimes e_j - e_j \otimes e_i) : 1 \leq i < j \leq m \right\},$$

which will be called *the matrix completion basis*. Here and in what follows  $\otimes$  denotes the tensor product of vectors or matrices. Note that, for a Hermitian matrix  $\rho$ , observing inner products  $\langle \rho, E_i \rangle$  with randomly picked matrices  $E_i$  from the above basis provides information about real and imaginary parts of the entries of the matrix, which explains the connection to the matrix completion problems. Another option is to consider the following basis of the space of all Hermitian matrices:

$$\left\{ e_i \otimes e_i : i = 1, \dots, m \right\} \cup \left\{ \frac{1}{2}(e_i \otimes e_j + e_j \otimes e_i) + \frac{i}{2}(e_i \otimes e_j - e_j \otimes e_i) : 1 \leq i < j \leq m \right\}.$$

Inner products of a Hermitian matrix  $\rho$  with the matrices of this basis are precisely the entries  $\rho_{ij}, i \leq j$  of matrix  $\rho$ . If now  $\Pi$  is the probability distribution (non-uniform) that assigns probabilities  $m^{-2}$  to the matrices  $e_i \otimes e_i$  corresponding to the diagonal entries and probabilities  $2m^{-2}$  to other matrices of the basis, then  $\mathbb{E}[\langle A, X \rangle]^2 = m^{-2} \|A\|_2^2$ . Sampling from this distribution is equivalent to sampling the entries of the matrix  $\rho$  at random (again, recall that when one learns an entry  $\rho_{ij}$  one also learns  $\rho_{ji} = \bar{\rho}_{ij}$ ).

Another example was studied by Gross et al (2009) and by Gross (2009). It is more directly related to the problems of quantum state tomography.

**Example 2. Pauli basis.** Let  $m = 2^k$ . Consider the *Pauli basis* in the space of  $2 \times 2$  matrices  $\mathbb{M}_2(\mathbb{C})$ :  $W_i := \frac{1}{\sqrt{2}}\sigma_i$ , where

$$\sigma_1 := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 := \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad \sigma_4 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

are the *Pauli matrices*. Note that the Pauli matrices are both Hermitian and unitary. The Pauli basis in  $\mathbb{M}_2(\mathbb{C})$  can be extended to a basis in the space of  $m \times m$  matrices  $\mathbb{M}_m(\mathbb{C})$ . These matrices define linear transformations acting in the linear space  $\mathbb{C}^m = \mathbb{C}^{2^k}$  that can be viewed as a  $k$ -fold tensor product of spaces  $\mathbb{C}^2 : \mathbb{C}^{2^k} = (\mathbb{C}^2)^{\otimes k}$ . Then, the Pauli basis in

the space of matrices  $\mathbb{M}_{2^k}(\mathbb{C})$  consists of all tensor products  $W_{i_1} \otimes \cdots \otimes W_{i_k}$ ,  $(i_1, \dots, i_k) \in \{1, 2, 3, 4\}^k$ . As before,  $X_1, \dots, X_n$  are i.i.d. random variables sampled from this set of tensor products. Essentially, this is a standard measurement model for a  $k$  qubit system frequently used in quantum information, in particular, in quantum state and quantum process tomography (see Nielsen and Chuang (2000), section 8.4.2).

**Example 3. Subgaussian design.** Another interesting class of examples includes *subgaussian design matrices*  $X$  such that  $\langle A, X \rangle$  is a subgaussian random variable for each  $A \in \mathbb{M}_m(\mathbb{C})$ . (Recall that a random variable  $\eta$  is called subgaussian with parameter  $\sigma$  iff, for all  $\lambda \in \mathbb{R}$ ,  $\mathbb{E}e^{\lambda\eta} \leq e^{\lambda^2\sigma^2/2}$ ). These examples are, probably, of less interest in applications to quantum state tomography, but this is an important model, closely related to randomized designs in compressed sensing, for which one can use powerful tools developed in the high-dimensional probability. For instance, one can consider the *Gaussian design*, where  $X$  is a symmetric random matrix with real entries such that  $\{X_{ij} : 1 \leq i \leq j \leq m\}$  are independent centered normal random variables with  $\mathbb{E}X_{ii}^2 = 1$ ,  $i = 1, \dots, m$  and  $\mathbb{E}X_{ij}^2 = \frac{1}{2}$ ,  $i < j$ . Alternatively, one can consider the *Rademacher design* assuming that  $X_{ii} = \varepsilon_{ii}$ ,  $i = 1, \dots, m$  and  $X_{ij} = \frac{1}{\sqrt{2}}\varepsilon_{ij}$ ,  $i < j$ , where  $\{\varepsilon_{ij} : 1 \leq i \leq j \leq m\}$  are i.i.d. Rademacher random variables (that is, random variables taking values  $+1$  or  $-1$  with probability  $1/2$  each). In both cases,  $\mathbb{E}|\langle A, X \rangle|^2 = \|A\|_2^2$ ,  $A \in \mathbb{M}_m(\mathbb{C})$  (such random matrices  $X$  will be called *isotropic*) and  $\langle A, X \rangle$  is a subgaussian random variable whose subgaussian parameter is equal to  $\|A\|_2$  (up to a constant).

The problems of this nature belong to a rapidly growing area of low rank matrix recovery. The most popular methods developed so far are based on nuclear norm regularization.

In what follows, the Euclidean norm in the space  $\mathbb{C}^m$  will be denoted by  $|\cdot|$  and the inner product will be denoted by  $\langle \cdot, \cdot \rangle$  (with a little abuse of notation since it has been already used for the Hilbert–Schmidt inner product between matrices). We will denote by  $\|\cdot\|_p, p \geq 1$  the *Schatten  $p$ -norm* of matrices in  $\mathbb{M}_m(\mathbb{C})$  (and, if needed, in other matrix spaces). Specifically,  $\|A\|_p := \left( \sum_{k=1}^m \lambda_k^p(|A|) \right)^{1/p}$ , where  $|A| := (A^*A)^{1/2}$  and, for a Hermitian matrix  $B$ ,  $\lambda_k(B), k = 1, \dots, m$  are the eigenvalues of  $B$  (usually arranged in the decreasing order). In particular,  $\|\cdot\|_1$  is the usual nuclear norm and  $\|\cdot\|_2$  is the Hilbert–Schmidt norm. We will use the notation  $\|\cdot\|$  for the operator norm. In addition to the metrics generated by these norms, some other distances will be of interest in connection to the statistical problems discussed in this paper. In particular, denoting

by  $\Pi$  the distribution of the design matrix  $X$ , we will write

$$\|A\|_{L_2(\Pi)}^2 := \int \langle A, x \rangle^2 \Pi(dx) = \mathbb{E} \langle A, X \rangle^2, \quad A \in \mathbb{M}_m(\mathbb{C})$$

and we will often use the corresponding  $L_2(\Pi)$ -distance between matrices (say, between two states  $S_1, S_2 \in \mathcal{S}$ ). This distance represents the prediction error in statistical problems in question.

In the noiseless case (i.e., when  $\xi_j \equiv 0$ ), the following estimator of  $\rho$  has been extensively studied, especially, in the case of matrix completion problems (see Candes and Recht (2009), Candes and Tao (2009), Gross (2009), Recht (2009) and references therein):

$$\hat{\rho} := \operatorname{argmin} \left\{ \|S\|_1 : S \in \mathbb{M}_m(\mathbb{C}), \langle S, X_j \rangle = Y_j, j = 1, \dots, n \right\}.$$

Under some assumptions that resemble the restricted isometry conditions used in compressed sensing, it was shown that, with a high probability,  $\hat{\rho} = \rho$  provided that the number  $n$  of observations is sufficiently large. Namely, up to logarithmic factors and constants, it should be of the order  $mr$ , where  $r$  is the rank of the target matrix  $\rho$ .

In the noisy case, one has to deal with a matrix regression problem and the following penalized least squares estimator, which is akin to the LASSO used in sparse regression, was proposed and studied (see, e.g., Candes and Plan (2009), Rohde and Tsybakov (2009)):

$$\hat{\rho}^\varepsilon := \operatorname{argmin}_{S \in \mathbb{M}_m(\mathbb{C})} \left[ n^{-1} \sum_{j=1}^n (Y_j - \operatorname{tr}(SX_j))^2 + \varepsilon \|S\|_1 \right], \quad (1.1)$$

where  $\varepsilon$  is a regularization parameter. Note that these estimators are not constrained to the set  $\mathcal{S}$  of density matrices (since for these matrices the nuclear norm is equal to 1). Candes and Plan (2009) have also studied another estimator based on the nuclear norm minimization subject to linear constraints that resembles the Dantzig selector used in compressed sensing and Rohde and Tsybakov (2009) suggested estimators based on nonconvex penalties involving Schatten “ $p$ -norms” for  $p < 1$ .

We will study the following estimator of the unknown state  $\rho$  defined as a solution of a penalized empirical risk minimization problem:

$$\hat{\rho}^\varepsilon := \operatorname{argmin}_{S \in \mathcal{S}} \left[ n^{-1} \sum_{j=1}^n (Y_j - \operatorname{tr}(SX_j))^2 + \varepsilon \operatorname{tr}(S \log S) \right], \quad (1.2)$$

where  $\varepsilon > 0$  is a regularization parameter. The penalty term is based on the functional  $\operatorname{tr}(S \log S) = -\mathcal{E}(S)$ , where  $\mathcal{E}(S)$  is the *von Neumann entropy* of state  $S$ . Thus, the

method considered in this paper is based on a trade-off between fitting the model by the least squares in the class of all density matrices and maximizing the entropy of the state.

One can also consider a slightly different estimator defined as follows:

$$\check{\rho}^\varepsilon := \operatorname{argmin}_{S \in \mathcal{S}} \left[ \int \langle S, x \rangle^2 \Pi(dx) - \frac{2}{n} \sum_{j=1}^n Y_j \operatorname{tr}(S X_j) + \varepsilon \operatorname{tr}(S \log S) \right]. \quad (1.3)$$

Of course, the estimator (1.3) requires the knowledge of the design distribution  $\Pi$  while the estimator (1.2) can be also used in the cases when  $\Pi$  is unknown. It happens that it is somewhat easier to study the properties of estimator (1.3) than of (1.2) for which one has to deal with more complicated empirical processes. Note that both optimization problems (1.2) and (1.3) are convex (this is based on convexity of the penalty term that follows from the concavity of von Neumann entropy, see Nielsen and Chuang (2000)). In what follows, we will study only the estimators defined by (1.2).

A commutative version of entropy penalization and its connections to sparse recovery problems in convex hulls of finite dictionaries have been studied by Koltchinskii (2009). In the current paper, this approach is extended to the noncommutative case.

## 2 An Overview of Main Results

The results of this paper include oracle inequalities for the  $L_2(\Pi)$ -error of the empirical solution  $\hat{\rho}^\varepsilon$ . They will be stated in a general form in sections 5 and 6. Here we formulate our results only in two of the special examples outlined in the Introduction: subgaussian isotropic design (such as Gaussian or Rademacher) and random sampling from the Pauli basis. Assume, for simplicity, that the noise  $\{\xi_j\}$  is a sequence of i.i.d.  $N(0, \sigma_\xi^2)$  random variables (i.e., it is a Gaussian noise).

Let  $t > 0$  be fixed and denote  $t_m := t + \log(2m)$ ,  $\tau_n := t + \log \log_2(2n)$ .

First we consider the case of subgaussian isotropic design. Note that in this case  $\|A\|_{L_2(\Pi)} = \|A\|_2$ ,  $A \in \mathbb{M}_m(\mathbb{C})$ . Given a subspace  $L \subset \mathbb{C}^m$ ,  $P_L$  denotes the orthogonal projection on  $L$  and  $L^\perp$  denotes its orthogonal complement.

**Theorem 1** *Suppose  $X$  is a subgaussian isotropic matrix. There exist constants  $C > 0, c > 0$  such that the following holds. Under the assumption that  $\tau_n \leq cn$ , for all  $\varepsilon \in$*

$[0, 1]$ , with probability at least  $1 - e^{-t}$

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq C \left( \varepsilon \left( \|\log \rho\| \wedge \log \frac{m}{\varepsilon} \right) \vee \sigma_\xi \sqrt{\frac{mt_m}{n}} \vee \right. \\ &\quad \left. (\sigma_\xi \vee \sqrt{m}) \frac{\sqrt{m}(\tau_n \log n \vee t_m)}{n} \right). \end{aligned} \quad (2.1)$$

Moreover, there exists a constant  $D > 0$  such that, for all  $\varepsilon \geq D\sigma_\xi \left( \sqrt{\frac{mt_m}{n}} \vee \frac{\sqrt{mt_m}}{n} \right)$ , with probability at least  $1 - e^{-t}$ ,

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq \inf_{S \in \mathcal{S}, L \subset \mathbb{C}^m} \left[ 2\|S - \rho\|_{L_2(\Pi)}^2 + C \left( \varepsilon^2 \|\log S\|_2^2 \vee \sigma_\xi^2 \frac{m \dim(L) + \tau_n}{n} \vee \right. \right. \\ &\quad \left. \left. \sigma_\xi \|P_{L^\perp} S P_{L^\perp}\|_1 \sqrt{\frac{mt_m}{n}} \vee (\sigma_\xi \vee \sqrt{m}) \frac{\sqrt{m}(\tau_n \log n \vee t_m)}{n} \right) \right]. \end{aligned} \quad (2.2)$$

This theorem includes two bounds on the  $L_2(\Pi)$ -error of  $\hat{\rho}^\varepsilon$ . The first bound (2.1) holds for all  $\varepsilon$  including  $\varepsilon = 0$ , which is the case of the unpenalized least squares estimator. The term  $\varepsilon \left( \|\log \rho\| \wedge \log \frac{m}{\varepsilon} \right)$  in this bound depends on the operator norm of  $\log \rho$  and it has to do with the approximation error of the entropy penalization method (see Section 4). The second bound (2.2) is an oracle inequality that controls the squared  $L_2(\Pi)$ -error of the estimator  $\hat{\rho}^\varepsilon$  in terms of approximation errors of oracles  $S \in \mathcal{S}$ . The term  $\varepsilon^2 \|\log S\|_2^2$  in this bound is also related to the approximation error of the entropy penalization method discussed in Section 4. This term depends on the Hilbert-Schmidt norm of  $\log S$ . The dependence on  $\varepsilon$  is better than in the first bound, but bound (2.2) holds only for the values of regularization parameter above certain threshold. Clearly, in the second bound, the oracles  $S$  are to be of full rank (otherwise,  $\log S$  does not exist and the right hand side of the bound becomes infinite). The random errors in these bounds are also different. In the first bound, it is of the order  $n^{-1/2}$  (up to logarithmic factors). In the second bound, the error term depends on how well the oracle  $S$  is approximated by low rank matrices. If there exists a subspace  $L$  of small dimension  $\dim(L)$  such that  $\|P_{L^\perp} S P_{L^\perp}\|_1$  is small (say, of the order  $n^{-1/2}$ ), then the random part of the error in (2.2) is essentially controlled by  $\sigma_\xi^2 \frac{\dim(L)m}{n}$ .

It will be shown later in the paper how to derive from the bounds of Theorem 1 and more general bounds for oracles of full rank some other inequalities for low rank oracles. In particular, for subgaussian isotropic design and Gaussian noise, this approach yields the following result. To simplify its formulation, we will assume that, for some constant  $c > 0$ ,  $\tau_n \leq cn$  and  $t_m \leq n$ .

**Theorem 2** Suppose  $X$  is a subgaussian isotropic matrix. There exist a constant  $c > 0$  and, for all sufficiently large  $D > 0$ , a constant  $C > 0$  such that, for  $\varepsilon := D\sigma_\xi\sqrt{\frac{mt_m}{n}}$ , with probability at least  $1 - e^{-t}$ ,

$$\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \inf_{S \in \mathcal{S}} \left[ 2\|S - \rho\|_{L_2(\Pi)}^2 + C \left( \frac{\sigma_\xi^2 \text{rank}(S) m t_m \log^2(mn)}{n} \vee \frac{m(\tau_n \log n \vee t_m)}{n} \right) \right]. \quad (2.3)$$

A simple consequence of the first bound of Theorem 1 and the bound of Theorem 2 is the following inequality that holds with probability at least  $1 - e^{-t}$  and with some  $C > 0$  for  $\varepsilon := D\sigma_\xi\sqrt{\frac{mt_m}{n}}$ :

$$\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq C \left[ \left( \sigma_\xi \sqrt{\frac{mt_m}{n}} \log(mn) \wedge \frac{\sigma_\xi^2 \text{rank}(\rho) m t_m \log^2(mn)}{n} \right) \vee \frac{m(\tau_n \log n \vee t_m)}{n} \right].$$

Next we consider the case of sampling from the Pauli basis. In this case,  $\|A\|_{L_2(\Pi)} = m^{-1}\|A\|_2$ ,  $A \in \mathbb{M}_m(\mathbb{C})$ . As before, we fix  $t > 0$  and assume that  $t_m \leq n$ .

**Theorem 3** Suppose that  $X$  is sampled at random from the uniform distribution  $\Pi$  on the Pauli basis. Then, there exists a constant  $C > 0$  such that, for all  $\varepsilon \in [0, 1]$ , with probability at least  $1 - e^{-t}$ ,

$$\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq C \left[ \varepsilon \left( \|\log \rho\| \wedge \log \left( \frac{m}{\varepsilon} \right) \right) \vee (\sigma_\xi \vee m^{-1/2}) \sqrt{\frac{t_m}{nm}} \right]. \quad (2.4)$$

In addition, for all sufficiently large  $D > 0$ , there exists a constant  $C > 0$  such that, for

$$\varepsilon := D(\sigma_\xi m^{-1/2} \vee m^{-1}) \sqrt{\frac{t_m}{n}},$$

with probability at least  $1 - e^{-t}$ ,

$$\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \inf_{S \in \mathcal{S}} \left[ 2\|S - \rho\|_{L_2(\Pi)}^2 + C(\sigma_\xi^2 \vee m^{-1}) \frac{\text{rank}(S) m t_m \log^2(mn)}{n} \right]. \quad (2.5)$$

Similarly to the previous theorems, one can easily derive from Theorem 3 the following bound

$$\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq C \left[ (\sigma_\xi \vee m^{-1/2}) \sqrt{\frac{t_m}{nm}} \log(mn) \wedge (\sigma_\xi^2 \vee m^{-1}) \frac{\text{rank}(\rho) m t_m \log^2(mn)}{n} \right]$$

that holds with probability at least  $1 - e^{-t}$  and with some  $C > 0$  for  $\varepsilon = D(\sigma_\xi m^{-1/2} \vee m^{-1}) \sqrt{\frac{t_m}{n}}$ .



It is worth mentioning that the results of sections 4, 5 provide a way to bound the error of estimator  $\hat{\rho}^\varepsilon$  not only in the  $L_2(\Pi)$ -distance, but also in other statistically important distances such as noncommutative Kullback-Leibler, Hellinger and nuclear norm distance (see Section 3.1 for their definitions). For instance, under the assumptions of Theorem 1, the following bound for the symmetrized Kullback-Leibler distance holds with probability at least  $1 - e^{-t}$  :

$$K(\hat{\rho}^\varepsilon; \rho) \leq \frac{C}{\varepsilon} \inf_{L \subset \mathbb{C}^m} \left[ \varepsilon^2 \|\log \rho\|_2^2 \vee \sigma_\xi^2 \frac{m \dim(L) + \tau_n}{n} \vee \sigma_\xi \|P_{L^\perp} \rho P_{L^\perp}\|_1 \sqrt{\frac{mt_m}{n}} \vee (\sigma_\xi \vee \sqrt{m}) \frac{\sqrt{m}(\tau_n \log n \vee t_m)}{n} \right]. \quad (2.6)$$

In the case of sampling from Pauli basis (as in Theorem 3), it is easy to derive from Theorem 5 of Section 5 (using also some bounds from the proofs of Proposition 5 and Corollary 1) the following bound on the squared Hellinger distance between  $\hat{\rho}^\varepsilon$  and  $\rho$  :

$$H^2(\hat{\rho}^\varepsilon; \rho) \leq C(\sigma_\xi \vee m^{-1/2}) \frac{\text{rank}(\rho) \sqrt{mt_m} \log^2(mn)}{\sqrt{n}}$$

that holds with probability at least  $1 - e^{-t}$  for  $\varepsilon = D(\sigma_\xi m^{-1/2} \vee m^{-1}) \sqrt{\frac{t_m}{n}}$ .

It has been already mentioned that the first bounds of theorems 1 and 3 (bounds (2.1) and (2.4)) hold for all  $\varepsilon \geq 0$ , even in the case of unpenalized least squares estimator with  $\varepsilon = 0$ . The random error parts of these bounds are (up to logarithmic factors) of the order  $n^{-1/2}$  as  $n \rightarrow \infty$ . Bounds (2.2), (2.3) and (2.5) are based on more subtle analysis taking into account the ranks of the oracles  $S$  approximating the true density matrix  $\rho$ . In these bounds, the size of the  $L_2(\Pi)$ -error  $\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2$  is determined by a trade-off between the approximation error  $\|S - \rho\|_{L_2(\Pi)}^2$  of an oracle  $S$  and the random error. In the case of bounds (2.3) and (2.5), the last error is of the order  $\frac{\sigma_\xi^2 \text{rank}(S)m}{n}$  (up to logarithmic factors), and it depends on the rank of the oracle  $S$ . In particular, taking  $S = \rho$ , we can conclude that  $\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2$  is bounded by  $\frac{\sigma_\xi^2 \text{rank}(\rho)m}{n}$  (up to constants and logarithmic factors). This means that von Neumann entropy penalization mimics oracles that know precisely which low rank matrices approximate  $\rho$  well and can estimate  $\rho$  by estimating a “small” number of parameters needed to describe such oracles. This could be compared with recent results for nuclear norm penalization (Candes and Plan (2009), Rohde and Tsybakov (2009)). Depending on the values of  $\sigma_\xi, m, n$  and other characteristics of the problem more “rough” bounds (2.1) and (2.4) might become even sharper than more “subtle” bounds (2.2), (2.3) and (2.5) (see Rohde and Tsybakov

(2009) for a discussion of a similar phenomenon). Since the random error term in more “subtle” bounds is proportional to  $\sigma_\xi^2$  and in the “rough” bounds it is proportional to  $\sigma_\xi$ , the “rough” bounds become sharper for the values of standard deviation of the noise  $\sigma_\xi$  above a threshold that depends on  $n$  and  $m$ . Thus, the rate of convergence of the  $L_2(\Pi)$ -error to zero in a particular asymptotic scenario (when certain characteristics are large) is determined by the bounds of both types.

Theorems 1, 2, 3 and other results of a similar nature will follow as corollaries from more general oracle inequalities that we establish under broader assumptions on the design distributions and on the noise. To prove these results, we need several tools from the empirical processes and random matrices theory, such as noncommutative Bernstein type inequalities and generic chaining bounds for empirical processes. We will discuss these results in Section 3 (as well as some properties of noncommutative Kullback-Leibler, Hellinger and other distances between density matrices). We will then study approximation error bounds for the solution of von Neumann entropy penalized true risk minimization problem (Section 4) and, finally, in sections 5 and 6, derive main results of the paper concerning random error bounds for the empirical solution  $\hat{\rho}^\varepsilon$ . More precisely, we bound the squared  $L_2(\Pi)$ -distance  $\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2$  and symmetrized Kullback-Leibler distance  $K(\hat{\rho}^\varepsilon; S)$  from  $\hat{\rho}^\varepsilon$  to an arbitrary “oracle”  $S \in \mathcal{S}$  and derive oracle inequalities for the squared  $L_2(\Pi)$ -error  $\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2$  of the empirical solution  $\hat{\rho}^\varepsilon$ . These results are first established for oracles  $S$  of full rank and expressed in terms of certain characteristics of the operator  $\log S$  (which is, essentially, a subgradient of the von Neumann entropy penalty used in (1.2)). Using simple techniques discussed in Section 4, we then develop the bounds for low rank oracles  $S$  (such as the bounds of theorems 2 and 3) and also obtain oracle inequalities for so called “Gibbs oracles”. Note that the logarithmic factors involved in the bounds of theorems 2 and 3 (and in other results of this type discussed later in the paper), in particular, the factor  $\log^2(mn)$ , are related to the need to bound certain norms of  $\log S$  for special oracles  $S \in \mathcal{S}$  (as in Theorem 1). In the case of  $\|S\|_1$ -penalization,  $\log S$  should be replaced with a version of  $\text{sign}(S)$  and one can avoid some of the logarithmic factors in this case.

### 3 Preliminaries: Distances in $\mathcal{S}$ , Empirical Processes and Exponential Inequalities for Random Matrices

#### 3.1 Noncommutative Kullback-Leibler and other distances

We will use noncommutative extensions of classical distances between probability distributions such as Kullback-Leibler and Hellinger distances. These extensions are common in quantum information theory (see Nielsen and Chuang (2000)). In particular, we will use *Kullback-Leibler divergence* between two states  $S_1, S_2 \in \mathcal{S}$  defined as

$$K(S_1 \| S_2) := \mathbb{E}_{S_1}(\log S_1 - \log S_2) = \text{tr}(S_1(\log S_1 - \log S_2))$$

and its symmetrized version

$$K(S_1; S_2) := K(S_1 \| S_2) + K(S_2 \| S_1) = \text{tr}((S_1 - S_2)(\log S_1 - \log S_2)).$$

We will also use a noncommutative version of *Hellinger distance* defined as follows. For any two states  $S_1, S_2 \in \mathcal{S}$ , let  $F(S_1, S_2) := \text{tr} \sqrt{S_1^{1/2} S_2 S_1^{1/2}}$ . This quantity is called the *fidelity* of states  $S_1, S_2$  (see, e.g., Nielsen and Chuang (2000), p. 409). Then, a natural definition of the squared Hellinger distance is  $H^2(S_1, S_2) := 2(1 - F(S_1, S_2))$ . A remarkable property of this distance is that

$$H^2(S_1, S_2) = \sup H^2(\{p_i\}; \{q_i\}) = \sup \sum_i \left( \sqrt{p_i} - \sqrt{q_i} \right)^2,$$

where the supremum is taken over all POVMs  $\{E_i\}$  (positive operator valued measures) and  $p_i := \text{tr}(S_1 E_i), q_i := \text{tr}(S_2 E_i)$ . [In the discrete case, a positive operator valued measure is a set  $\{E_i\}$  of Hermitian nonnegatively definite matrices such that  $\sum_i E_i = I$ ]. Thus, the quantum Hellinger distance is just the largest “classical” Hellinger distance between the probability distributions  $\{p_i\}, \{q_i\}$  of a “measurement”  $\{E_i\}$  in the states  $S_1, S_2$  (see Nielsen and Chuang (2000), p. 412). The same property also holds for two other important “distances”, the trace distance  $\|S_1 - S_2\|_1$  and the Kullback-Leibler divergence  $K(S_1 \| S_2)$  (see, e.g., Klauck et al (2007)). These properties immediately imply an extension of classical inequalities for these distances:

$$\|S_1 - S_2\|_1^2 \leq H^2(S_1, S_2) \leq K(S_1 \| S_2).$$

They also imply the following simple proposition used below. It shows that, if two matrices  $S_1, S_2$  are close in the Hellinger distance and one of them (say,  $S_2$ ) is “approximately

low rank” in the sense that there exists a subspace  $L \subset \mathbb{C}^m$  of small dimension such that  $\|P_{L^\perp} S_2 P_{L^\perp}\|_1$  is small, then another matrix  $S_1$  is also “approximately low rank” with the same “support”  $L$ .

**Proposition 1** *For all subspaces  $L \subset \mathbb{C}^m$  and all  $S_1, S_2 \in \mathcal{S}$ ,*

$$\|P_L S_1 P_L\|_1 \leq 2\|P_L S_2 P_L\|_1 + 2H^2(S_1, S_2).$$

**Proof.** Indeed, take an orthonormal basis  $\{e_1, \dots, e_m\}$  in  $\mathbb{C}^m$  such that  $L = \text{l.s.}(\{e_1, \dots, e_k\})$ . Let  $p_j := \langle S_1 e_j, e_j \rangle = \text{tr}(S_1(e_j \otimes e_j))$  and  $q_j := \langle S_2 e_j, e_j \rangle = \text{tr}(S_2(e_j \otimes e_j))$ . Then

$$H^2(S_1, S_2) \geq \sum_{j=1}^m \left( \sqrt{p_j} - \sqrt{q_j} \right)^2 \geq \sum_{j=1}^k \left( \sqrt{p_j} - \sqrt{q_j} \right)^2 = \sum_{j=1}^k p_j + \sum_{j=1}^k q_j - 2 \sum_{j=1}^k \sqrt{p_j} \sqrt{q_j},$$

which implies (using that  $2\sqrt{ab} \leq a/2 + 2b$ )

$$\|P_L S_1 P_L\|_1 = \sum_{j=1}^k p_j \leq 2 \sum_{j=1}^k \sqrt{p_j} \sqrt{q_j} - \sum_{j=1}^k q_j + H^2(S_1, S_2) \leq$$

$$\frac{1}{2} \sum_{j=1}^k p_j + \sum_{j=1}^k q_j + H^2(S_1, S_2) = \frac{1}{2} \|P_L S_1 P_L\|_1 + \|P_L S_2 P_L\|_1 + H^2(S_1, S_2),$$

and the result follows. □

### 3.2 Empirical processes bounds

We will use several inequalities for empirical processes indexed by a class of measurable functions  $\mathcal{F}$  defined on an arbitrary measurable space  $(S, \mathcal{A})$ . Let  $X, X_1, \dots, X_n$  be i.i.d. random variables in  $(S, \mathcal{A})$  with common distribution  $P$ . If  $\mathcal{F}$  is uniformly bounded by a number  $U$ , then Bousquet’s version of the famous Talagrand’s concentration inequality for empirical processes implies that, for all  $t > 0$ , with probability at least  $1 - e^{-t}$

$$\sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^n f(X_j) - \mathbb{E} f(X) \right| \leq 2 \left[ \mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^n f(X_j) - \mathbb{E} f(X) \right| + \sigma \sqrt{\frac{t}{n}} + U \frac{t}{n} \right],$$

where  $\sigma^2 := \sup_{f \in \mathcal{F}} \text{Var}_P(f(X))$ . We will also need a version of this bound for function classes that are not necessarily uniformly bounded. Such a bound was recently proved by Adamczak (2008). Let  $F(x) \geq \sup_{f \in \mathcal{F}} |f(x)|, x \in S$ , be an envelope of the class. It

follows from Theorem 4 of Adamczak (2008) that, there exists a constant  $K > 0$  such that for all  $t > 0$  with probability at least  $1 - e^{-t}$

$$\sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^n f(X_j) - \mathbb{E}f(X) \right| \leq K \left[ \mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^n f(X_j) - \mathbb{E}f(X) \right| + \sigma \sqrt{\frac{t}{n}} + \left\| \max_{1 \leq j \leq n} |F(X_j)| \right\|_{\psi_1} \frac{t}{n} \right].$$

In addition to this, we will need to bound the following expectation:

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^n f^2(X_j) - \mathbb{E}f^2(X) \right|.$$

A usual approach to this problem is to use symmetrization inequality to replace the empirical process by a Rademacher process, and then to use Talagrand's comparison (contraction) inequality (see, e.g., Ledoux and Talagrand (1991), Section 4.5) to get rid of the squares. This, however, would require the class  $\mathcal{F}$  to be uniformly bounded by some  $U > 0$ , which is not too large. This approach is not sufficient in the case of subgaussian design considered in the last section. A more subtle approach has been developed in the recent years by Klartag and Mendelson (2005), Mendelson (2010) and it is based on generic chaining bounds.

Talagrand's *generic chaining complexity* (see Talagrand (2005)) of a metric space  $(T, d)$  is defined as follows. An admissible sequence  $\{\Delta_n\}_{n \geq 0}$  is an increasing sequence of partitions of  $T$  (i.e., each next partition is a refinement of the previous one) such that  $\text{card}(\Delta_0) = 1$  and  $\text{card}(\Delta_n) \leq 2^{2^n}$ ,  $n \geq 1$ . For  $t \in T$ ,  $\Delta_n(t)$  denotes the unique subset in  $\Delta_n$  that contains  $t$ . For a set  $A \subset T$ ,  $D(A)$  denotes its diameter. Then, define the generic chaining complexity  $\gamma_2(T; d)$  as

$$\gamma_2(T; d) := \inf_{\{\Delta_n\}_{n \geq 0}} \sup_{t \in T} \sum_{n \geq 0} 2^{n/2} D(\Delta_n(t)),$$

where the inf is taken over all admissible sequences of partitions.

If  $\{X(t) : t \in T\}$  is a centered Gaussian process such that  $\mathbb{E}(X(t) - X(s))^2 = d^2(t, s)$ ,  $t, s \in T$ , then it was proved by Talagrand that

$$K^{-1} \gamma_2(T; d) \leq \mathbb{E} \sup_{t \in T} X(t) \leq K \gamma_2(T; d),$$

where  $K > 0$  is a universal constant. Thus, the generic chaining complexity  $\gamma_2(T; d)$  is a natural characteristic of the size of the Gaussian process  $X(t)$ ,  $t \in T$ .

Similar quantities can be also used to control the size of empirical processes indexed by a function class  $\mathcal{F}$ . It is natural to define  $\gamma_2(\mathcal{F}; L_2(P))$ , that is,  $\gamma_2(\mathcal{F}; d)$ , where  $d$  is

the  $L_2(P)$ -distance. Some other distances are also useful, for instance, the  $\psi_2$ -distance associated with the probability space  $(S, \mathcal{A}, P)$ . Recall that, for a convex increasing function  $\psi$  with  $\psi(0) = 0$ ,

$$\|f\|_\psi := \inf \left\{ C > 0 : \int_S \psi \left( \frac{|f|}{C} \right) dP \leq 1 \right\}$$

(see van der Vaart and Wellner (1996), p. 95). If  $\psi(u) = u^p, u \geq 0$ , for some  $p \geq 1$ , the corresponding  $\psi$ -norm is just the  $L_p$ -norm. Other important choices are functions  $\psi_\alpha(t) = e^{t^\alpha} - 1, t \geq 0, \alpha \geq 1$ , especially,  $\psi_2$  that is related to subgaussian tails of  $f$  and  $\psi_1$  that is related to subexponential tails.

The generic chaining complexity that corresponds to the  $\psi_2$ -distance will be denoted by  $\gamma_2(\mathcal{F}; \psi_2)$ . Mendelson (2010) proved the following deep result (strengthening previous results by Klartag and Mendelson (2005)). Suppose that  $\mathcal{F}$  is a symmetric class, that is,  $f \in \mathcal{F}$  implies  $-f \in \mathcal{F}$ , and  $Pf = \mathbb{E}f(X) = 0, f \in \mathcal{F}$ . Then, for some universal constant  $K > 0$ ,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{j=1}^n f^2(X_j) - \mathbb{E} f^2(X) \right| \leq K \left[ \sup_{f \in \mathcal{F}} \|f\|_{\psi_1} \frac{\gamma_2(\mathcal{F}; \psi_2)}{\sqrt{n}} \vee \frac{\gamma_2^2(\mathcal{F}; \psi_2)}{n} \right].$$

### 3.3 Noncommutative Bernstein type inequalities

We will need the following operator version of Bernstein's inequality which is due to Ahlswede and Winter (2002) (and which has been already successfully used in the low rank recovery problems by Gross et al (2009), Gross (2009), Recht (2009)).

In this subsection, assume that  $X, X_1, \dots, X_n$  are i.i.d. random Hermitian  $m \times m$  matrices with  $\mathbb{E}X = 0$  and  $\sigma_X^2 := \|\mathbb{E}X^2\|$ .

**Bernstein's inequality for operator valued r.v.** *Suppose that  $\|X\| \leq U$  for some  $U > 0$ . Then*

$$\mathbb{P} \left\{ \|X_1 + \dots + X_n\| \geq t \right\} \leq 2m \exp \left\{ -\frac{t^2}{2\sigma_X^2 n + 2Ut/3} \right\}. \quad (3.1)$$

In fact, we will frequently use the following bound that immediately follows from the version of Bernstein's inequality given above: *for all  $t > 0$ , with probability at least  $1 - e^{-t}$*

$$\left\| \frac{X_1 + \dots + X_n}{n} \right\| \leq 2 \left( \sigma_X \sqrt{\frac{t + \log(2m)}{n}} \vee U \frac{t + \log(2m)}{n} \right). \quad (3.2)$$

Moreover, it is possible to replace the  $L_\infty$ -bound  $U$  on  $\|X\|$  in the above inequality by bounds on the weaker  $\psi_\alpha$ -norms. Denote  $U_X^{(\alpha)} := \left\| \|X\| \right\|_{\psi_\alpha}$ ,  $\alpha \geq 1$ .

**Proposition 2** *Let  $\alpha \geq 1$ . There exists a constant  $C > 0$  such that, for all  $t > 0$ , with probability at least  $1 - e^{-t}$*

$$\left\| \frac{X_1 + \cdots + X_n}{n} \right\| \leq C \left( \sigma_X \sqrt{\frac{t + \log(2m)}{n}} \vee U_X^{(\alpha)} \left( \log \frac{U_X^{(\alpha)}}{\sigma_X} \right)^{1/\alpha} \frac{t + \log(2m)}{n} \right). \quad (3.3)$$

Note that, in the limit  $\alpha \rightarrow \infty$ , inequality (3.3) coincides with (3.2) (up to a constant).

**Proof.** Similarly to the proof of (3.1) discussed in the literature (Ahlsvede and Winter (2002), Gross (2009), Recht (2009)), we follow the standard derivation of classical Bernstein's inequality and we use the well known *Golden-Thompson inequality* (see, e.g., Simon (1979), p. 94): for arbitrary Hermitian matrices  $A, B \in \mathbb{M}_m(\mathbb{C})$ ,  $\text{tr}(e^{A+B}) \leq \text{tr}(e^A e^B)$ . Let  $Y_n := X_1 + \cdots + X_n$ . Note that  $\|Y_n\| < t$  if and only if  $-tI_m < Y_n < tI_m$ . Therefore,

$$\mathbb{P}\{\|Y_n\| \geq t\} = \mathbb{P}\{Y_n \not\leq tI_m\} + \mathbb{P}\{Y_n \not\geq -tI_m\}. \quad (3.4)$$

The following bounds are straightforward by simple matrix algebra:

$$\mathbb{P}\{Y_n \not\leq tI_m\} = \mathbb{P}\{e^{\lambda Y_n} \not\leq e^{\lambda t I_m}\} \leq \mathbb{P}\left\{\text{tr}\left(e^{\lambda Y_n}\right) \geq e^{\lambda t}\right\} \leq e^{-\lambda t} \mathbb{E} \text{tr}(e^{\lambda Y_n}). \quad (3.5)$$

To bound the expected value in the right hand side, we use independence of random variables  $X_1, \dots, X_n$  and Golden-Thompson inequality:

$$\begin{aligned} \mathbb{E} \text{tr}(e^{\lambda Y_n}) &= \mathbb{E} \text{tr}(e^{\lambda Y_{n-1} + \lambda X_n}) \leq \mathbb{E} \text{tr}(e^{\lambda Y_{n-1}} e^{\lambda X_n}) = \text{tr}\left(\mathbb{E}\left(e^{\lambda Y_{n-1}} e^{\lambda X_n}\right)\right) = \\ &= \text{tr}\left(\mathbb{E} e^{\lambda Y_{n-1}} \mathbb{E} e^{\lambda X_n}\right) \leq \mathbb{E} \text{tr}(e^{\lambda Y_{n-1}}) \left\| \mathbb{E} e^{\lambda X_n} \right\|. \end{aligned}$$

By induction, we conclude that

$$\mathbb{E} \text{tr}(e^{\lambda Y_n}) \leq \mathbb{E} \text{tr}(e^{\lambda X_1}) \left\| \mathbb{E} e^{\lambda X_2} \right\| \cdots \left\| \mathbb{E} e^{\lambda X_n} \right\|.$$

Since  $\mathbb{E} \text{tr}(e^{\lambda X_1}) = \text{tr}(\mathbb{E} e^{\lambda X_1}) \leq m \left\| \mathbb{E} e^{\lambda X} \right\|$ , we get

$$\mathbb{E} \text{tr}(e^{\lambda Y_n}) \leq m \left\| \mathbb{E} e^{\lambda X} \right\|^n. \quad (3.6)$$

It remains to bound the norm  $\|\mathbb{E}e^{\lambda X}\|$ . To this end, we use Taylor expansion and the condition  $\mathbb{E}X = 0$  to get

$$\begin{aligned}\mathbb{E}e^{\lambda X} &= I_m + \mathbb{E}\lambda^2 X^2 \left[ \frac{1}{2!} + \frac{\lambda X}{3!} + \frac{\lambda^2 X^2}{4!} + \dots \right] \leq \\ I_m + \lambda^2 \mathbb{E}X^2 \left[ \frac{1}{2!} + \frac{\lambda \|X\|}{3!} + \frac{\lambda^2 \|X\|^2}{4!} + \dots \right] &= I_m + \lambda^2 \mathbb{E}X^2 \left[ \frac{e^{\lambda \|X\|} - 1 - \lambda \|X\|}{\lambda^2 \|X\|^2} \right].\end{aligned}$$

Therefore, for all  $\tau > 0$ ,

$$\begin{aligned}\|\mathbb{E}e^{\lambda X}\| &\leq 1 + \lambda^2 \left\| \mathbb{E}X^2 \left[ \frac{e^{\lambda \|X\|} - 1 - \lambda \|X\|}{\lambda^2 \|X\|^2} \right] \right\| \leq \\ 1 + \lambda^2 \left\| \mathbb{E}X^2 \right\| \left[ \frac{e^{\lambda \tau} - 1 - \lambda \tau}{\lambda^2 \tau^2} \right] &+ \lambda^2 \mathbb{E}\|X\|^2 \left[ \frac{e^{\lambda \|X\|} - 1 - \lambda \|X\|}{\lambda^2 \|X\|^2} \right] I(\|X\| \geq \tau).\end{aligned}$$

Let  $M := 2(\log 2)^{1/\alpha} U_X^{(\alpha)}$  and assume that  $\lambda \leq 1/M$ . Then

$$\begin{aligned}\mathbb{E}\|X\|^2 \left[ \frac{e^{\lambda \|X\|} - 1 - \lambda \|X\|}{\lambda^2 \|X\|^2} \right] I(\|X\| \geq \tau) &\leq M^2 \mathbb{E}e^{\|X\|/M} I(\|X\| \geq \tau) \leq \\ M^2 \mathbb{E}^{1/2} e^{2\|X\|/M} \mathbb{P}^{1/2} \{\|X\| \geq \tau\}.\end{aligned}$$

Since, for  $\alpha \geq 1$ ,  $M = 2(\log 2)^{1/\alpha} \left\| \|X\| \right\|_{\psi_\alpha} \geq 2 \left\| \|X\| \right\|_{\psi_1}$  (see van der Vaart and Wellner (1996), p. 95), we have  $\mathbb{E}e^{2\|X\|/M} \leq 2$  and also

$$\mathbb{P}\{\|X\| \geq \tau\} \leq \exp \left\{ -2^\alpha \log 2 \left( \frac{\tau}{M} \right)^\alpha \right\}.$$

As a result, we get the following bound

$$\|\mathbb{E}e^{\lambda X}\| \leq 1 + \lambda^2 \sigma_X^2 \left[ \frac{e^{\lambda \tau} - 1 - \lambda \tau}{\lambda^2 \tau^2} \right] + 2^{1/2} \lambda^2 M^2 \exp \left\{ -2^{\alpha-1} \log 2 \left( \frac{\tau}{M} \right)^\alpha \right\}.$$

Let  $\tau := M \frac{2^{1/\alpha-1}}{(\log 2)^{1/\alpha}} \log^{1/\alpha} \frac{M^2}{\sigma_X^2}$  and suppose that  $\lambda$  satisfies the condition  $\lambda \tau \leq 1$ . Then, the following bound holds with some constant  $C_1 > 0$ :

$$\|\mathbb{E}e^{\lambda X}\| \leq 1 + C_1 \lambda^2 \sigma_X^2 \leq \exp \{ C_1 \lambda^2 \sigma_X^2 \}.$$

Thus, we proved that there exist constants  $C_1, C_2 > 0$  such that, for all  $\lambda$  satisfying the condition

$$\lambda U_X^{(\alpha)} \left( \log \frac{U_X^{(\alpha)}}{\sigma_X} \right)^{1/\alpha} \leq C_2, \quad (3.7)$$



we have  $\|\mathbb{E}e^{\lambda X}\| \leq \exp\{C_1\lambda^2\sigma_X^2\}$ . This can be combined with (3.4), (3.5) and (3.6) to get

$$\mathbb{P}\{\|Y_n\| \geq t\} \leq 2m \exp\left\{-\lambda t + C_1\lambda^2 n \sigma_X^2\right\}.$$

It remains now to minimize the last bound with respect to all  $\lambda$  satisfying (3.7) to get that, for some constant  $K > 0$ ,

$$\mathbb{P}\{\|Y_n\| \geq t\} \leq 2m \exp\left\{-\frac{1}{K} \frac{t^2}{n\sigma_X^2 + tU_X^{(\alpha)} \log^{1/\alpha}(U_X^{(\alpha)}/\sigma_X)}\right\},$$

which immediately implies (3.3).  $\square$

Note that, in a standard way, one can deduce bounds on the expectation from the exponential bounds on tail probabilities. In particular, (3.1) implies that

$$\mathbb{E}\left\|\frac{X_1 + \dots + X_n}{n}\right\| \leq C\left(\sigma_X \sqrt{\frac{\log(2m)}{n}} \bigvee U \frac{\log(2m)}{n}\right). \quad (3.8)$$

Similarly, Proposition 2 implies that

$$\mathbb{E}\left\|\frac{X_1 + \dots + X_n}{n}\right\| \leq C\left(\sigma_X \sqrt{\frac{\log(2m)}{n}} \bigvee U_X^{(\alpha)} \left(\log \frac{U_X^{(\alpha)}}{\sigma_X}\right)^{1/\alpha} \frac{\log(2m)}{n}\right) \quad (3.9)$$

Combining the last bounds with Talagrand's concentration inequality leads to somewhat different versions of bounds (3.2) and (3.3) that can be better in some applications. Namely, denote

$$\tilde{\sigma}_X^2 := \sup_{u,v \in \mathbb{C}^m, |u| \leq 1, |v| \leq 1} \mathbb{E}|\langle Xu, v \rangle|^2.$$

It is easy to check that  $\tilde{\sigma}_X^2 \leq \sigma_X^2$ . Moreover, in some cases, it can be significantly smaller (for instance, if  $X$  is sampled at random from the matrix completion basis, then  $\sigma_X^2$  is of the order  $m^{-1}$  and  $\tilde{\sigma}_X^2$  is equal to  $m^{-2}$ ). The expectation bound (3.8) and Talagrand's concentration inequality imply that with probability at least  $1 - e^{-t}$

$$\left\|\frac{X_1 + \dots + X_n}{n}\right\| \leq C\left(\sigma_X \sqrt{\frac{\log(2m)}{n}} \bigvee \tilde{\sigma}_X \sqrt{\frac{t}{n}} \bigvee U \frac{\log(2m)}{n} \bigvee U \frac{t}{n}\right). \quad (3.10)$$

Similarly, combining the expectation bound (3.9) for  $\alpha = 1$  with Adamczak's version of Talagrand's inequality (see Section 3.2), we get that with probability at least  $1 - e^{-t}$

$$\left\|\frac{X_1 + \dots + X_n}{n}\right\| \leq C\left(\sigma_X \sqrt{\frac{\log(2m)}{n}} \bigvee \tilde{\sigma}_X \sqrt{\frac{t}{n}} \bigvee U_X^{(1)} \left(\log \frac{U_X^{(1)}}{\sigma_X}\right) \frac{\log(2m)}{n} \bigvee U_X^{(1)} \frac{t \log n}{n}\right). \quad (3.11)$$

In the examples when  $\tilde{\sigma}_X^2$  is significantly smaller than  $\sigma_X^2$ , these bounds might be better than (3.2) and (3.3), especially, when they are used for large values of  $t$ .

In principle, using bounds (3.10) and (3.11) in the proofs of the following sections instead of (3.2) and (3.3) provides a way to obtain probabilistic oracle inequalities with probabilities of the error decreasing exponentially with  $m$  or  $n$  (this is the way in which error bounds are written in the papers by Candes and Plan (2009) and Rohde and Tsybakov (2009)). We are not pursuing this approach here.

## 4 Approximation Error

A natural first step in the analysis of the problem is to study its version with the true risk instead of the empirical risk. The true risk with respect to the quadratic loss is equal to

$$\mathbb{E}(Y - \langle S, X \rangle)^2 = \mathbb{E}(\langle \rho, X \rangle + \xi - \langle S, X \rangle)^2 = \mathbb{E}\langle S - \rho, X \rangle^2 + \mathbb{E}\xi^2,$$

where we used the assumptions that  $X$  and  $\xi$  are independent and  $\mathbb{E}\xi = 0$ . Thus, the penalized true risk minimization problem becomes

$$\rho^\varepsilon := \operatorname{argmin}_{S \in \mathcal{S}} \left[ \mathbb{E}\langle S - \rho, X \rangle^2 + \varepsilon \operatorname{tr}(S \log S) \right] \quad (4.1)$$

and the goal is to study the error of approximation of  $\rho$  by  $\rho^\varepsilon$  depending on the value of regularization parameter  $\varepsilon > 0$ . The next propositions show that if there exists an oracle  $S \in \mathcal{S}$  that provides a good approximation of the true state  $\rho$  in a sense that  $\|S - \rho\|_{L_2(\Pi)}$  is small, then  $\rho^\varepsilon$  belongs to an  $L_2(\Pi)$ -ball around  $S$  of small enough radius that can be controlled in terms of the operator norm  $\|\log S\|$  or in terms of more subtle characteristics of the oracle  $S$ . They also provide upper bounds on the approximation error  $\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}$ .

We will first obtain a simple bound on  $\|\rho^\varepsilon - S\|_{L_2(\Pi)}$  for an arbitrary oracle  $S \in \mathcal{S}$  of full rank expressed in terms of the operator norm  $\|\log S\|$  of its logarithm. For simplicity, we assume that  $\|\log S\| = +\infty$  in the case when  $\operatorname{rank}(S) < m$  (and  $\log S$  is not defined). Note, however, that  $\operatorname{tr}(S \log S)$  is well defined and finite even in the case when  $\operatorname{rank}(S) < m$ .

**Proposition 3** *For all  $S \in \mathcal{S}$ ,  $\|\rho^\varepsilon - S\|_{L_2(\Pi)} \leq \|S - \rho\|_{L_2(\Pi)} + \sqrt{\varepsilon \|\log S\|}$ . This implies that*

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)} \leq 2\|S - \rho\|_{L_2(\Pi)} + \sqrt{\varepsilon \|\log S\|},$$

*and, in particular, for  $S = \rho$ ,  $\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \varepsilon \|\log \rho\|$ .*

For a differentiable mapping  $g$  from an open subset  $G \subset \mathbb{M}_m(\mathbb{C})$  into  $\mathbb{M}_m(\mathbb{C})$ , denote by  $Dg(A; H)$  its differential at a matrix  $A \in G$  in the direction  $H \in \mathbb{M}_m(\mathbb{C})$ , that is,  $Dg(A; H)$  is linear with respect to  $H$  and

$$g(A + H) = g(A) + Dg(A; H) + o(\|H\|) \text{ as } \|H\| \rightarrow 0.$$

The following lemma is a simple corollary of Theorem V.3.3 in Bhatia (1996):

**Lemma 1** *Let  $f$  be a function continuously differentiable in an open interval  $I \subset \mathbb{R}$ . Suppose that  $A$  is a Hermitian matrix whose spectrum belongs to  $I$ . Then the mapping  $B \mapsto g(B) := \text{tr}(f(B))$  is differentiable at  $A$  and  $Dg(A; H) = \text{tr}(f'(A)H)$ .*

**Proof of Proposition 3.** Denote the penalized risk

$$L(S) := \mathbb{E}\langle S - \rho, X \rangle^2 + \varepsilon \text{tr}(S \log S).$$

It is easy to see that the solution  $\rho^\varepsilon$  of problem (4.1) is a full rank matrix. To prove this, assume that  $\text{rank}(\rho^\varepsilon) < m$ . Let  $\tilde{\rho} := (1 - \delta)\rho^\varepsilon + \delta I_m$ , where  $I_m$  is the  $m \times m$  identity matrix. Then, for small enough  $\delta$ ,  $\tilde{\rho}$  is a full rank matrix and it is straightforward to show that the penalized risk  $L(\tilde{\rho})$  is strictly smaller than  $L(\rho^\varepsilon)$  (for some small  $\delta > 0$ ). It is also easy to check that, for any  $S \in \mathcal{S}$  of full rank,  $\log S$  is well defined and the differential of the functional  $L$  in the direction  $\nu \in \mathbb{M}_m(\mathbb{C})$  is equal to

$$DL(S; \nu) = 2\mathbb{E}\langle S - \rho, X \rangle \langle \nu, X \rangle + \varepsilon \text{tr}(\nu \log S).$$

This follows from the fact that the first term of the functional  $L$  is differentiable since it is quadratic. The differentiability of the penalty term is based on Lemma 1 (it is enough to apply this lemma to the function  $f(u) = u \log u$ ). Since  $\rho^\varepsilon$  is the minimal point of  $L$  in  $\mathcal{S}$ , we can conclude that, for an arbitrary  $S \in \mathcal{S}$ ,  $DL(\rho^\varepsilon; S - \rho^\varepsilon) \geq 0$ . This implies that

$$DL(S; S - \rho^\varepsilon) - DL(\rho^\varepsilon; S - \rho^\varepsilon) \leq DL(S; S - \rho^\varepsilon),$$

which, by a simple algebra, becomes

$$2\|S - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \varepsilon K(S; \rho^\varepsilon) \leq 2\mathbb{E}\langle S - \rho, X \rangle \langle S - \rho^\varepsilon, X \rangle + \varepsilon \langle S - \rho^\varepsilon, \log S \rangle. \quad (4.2)$$

To conclude the proof, note that (4.2), the bound  $\|S - \rho^\varepsilon\|_1 \leq 2$  and Cauchy-Schwarz inequality imply that

$$2\|S - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \varepsilon K(S; \rho^\varepsilon) \leq 2\|S - \rho\|_{L_2(\Pi)}\|S - \rho^\varepsilon\|_{L_2(\Pi)} + 2\varepsilon\|\log S\|.$$

Solving the last inequality with respect to  $\|\rho^\varepsilon - S\|_{L_2(\Pi)}$  and using the fact that  $K(S; \rho^\varepsilon) \geq 0$ , yields the bound

$$\|\rho^\varepsilon - S\|_{L_2(\Pi)} \leq \frac{\|S - \rho\|_{L_2(\Pi)}}{2} + \sqrt{\frac{\|S - \rho\|_{L_2(\Pi)}^2}{4} + \varepsilon \|\log S\|},$$

which implies  $\|\rho^\varepsilon - S\|_{L_2(\Pi)} \leq \|S - \rho\|_{L_2(\Pi)} + \sqrt{\varepsilon \|\log S\|}$ , and the result follows.  $\square$

To obtain more subtle bounds with approximation error of the order  $O(\varepsilon^2)$  instead of  $O(\varepsilon)$ , we introduce and use the following quantity

$$a(W) := a_\Pi(W) := a_X(W) := \sup \left\{ \langle W, U \rangle : U \in \mathbb{M}_m(\mathbb{C}), U = U^*, \text{tr}(U) = 0, \|U\|_{L_2(\Pi)} = 1 \right\},$$

which will be called the *alignment coefficient* of  $W$ . Similar quantities were used in the commutative case (Koltchinskii (2009)). Note that, for all constants  $c$ ,

$$a(W + cI_m) = a(W) \quad (4.3)$$

(since  $\langle I_m, U \rangle = 0$  for all  $U$  of zero trace). In addition, we have

$$a_{cX}(W) = \frac{1}{|c|} a_X(W), \quad c \neq 0. \quad (4.4)$$

Let  $\{E_i : i = 1, \dots, m^2\}$  be an orthonormal basis of  $\mathbb{M}_m(\mathbb{C})$  consisting of Hermitian matrices and let  $\mathcal{K} := \left( \langle E_j, E_k \rangle_{L_2(\Pi)} \right)_{j,k=1}^{m^2}$  be the Gram matrix of the functions  $\{\langle E_j, \cdot \rangle : j = 1, \dots, m^2\}$  in the space  $L_2(\Pi)$ . Clearly, the mapping  $J : \mathbb{M}_m(\mathbb{C}) \mapsto \ell_2^{m^2}(\mathbb{C})$ ,

$$JU = \left( \langle U, E_j \rangle : j = 1, \dots, m^2 \right), \quad U \in \mathbb{M}_m(\mathbb{C}),$$

is an isometry. If now we define  $\bar{\mathcal{K}} : \mathbb{M}_m(\mathbb{C}) \mapsto \mathbb{M}_m(\mathbb{C})$  as  $\bar{\mathcal{K}} := J^{-1}\mathcal{K}J$ , then we also have  $\bar{\mathcal{K}}^{1/2} = J^{-1}\mathcal{K}^{1/2}J$ ,  $\bar{\mathcal{K}}^{-1/2} = J^{-1}\mathcal{K}^{-1/2}J$ . As a consequence, for any matrix  $U = \sum_{j=1}^{m^2} u_j E_j$ ,

$$\|U\|_{L_2(\Pi)}^2 = \sum_{j,k=1}^{m^2} \langle E_j, E_k \rangle_{L_2(\Pi)} u_j \bar{u}_k = \langle \mathcal{K}u, u \rangle_{\ell_2} = \|\mathcal{K}^{1/2}u\|_{\ell_2}^2 = \|\bar{\mathcal{K}}^{1/2}U\|_2^2,$$

and it is not hard to conclude that  $a(W) \leq \|\bar{\mathcal{K}}^{-1/2}W\|_2$ . Moreover, in view of (4.3), for an arbitrary scalar  $c$ ,

$$a(W) \leq \|\bar{\mathcal{K}}^{-1/2}(W + cI_m)\|_2.$$

This shows that the size of  $a(W)$  depends on how  $W$  is “aligned” with the eigenspaces of the Gram matrix  $\mathcal{K}$ . In a special case when, for all  $A$ ,  $\|A\|_{L_2(\Pi)} = \|A\|_2$ , the functions  $\{\langle E_j, \cdot \rangle : j = 1, \dots, m^2\}$  form an orthonormal system in the space  $L_2(\Pi)$  and the Gram matrix  $\mathcal{K}$  is the identity matrix. In this case, we simply have the bound

$$a(W) \leq \inf_c \|W + cI_m\|_2.$$

In the next statement, we use the alignment coefficient  $a(\log S)$  to control the  $L_2(\Pi)$ -distance  $\|\rho^\varepsilon - S\|_{L_2(\Pi)}$  and the Kullback-Leibler “distance”  $K(\rho^\varepsilon; S)$  from the true solution  $\rho^\varepsilon$  to an arbitrary oracle  $S$ .

**Proposition 4** *For all  $S \in \mathcal{S}$ ,*

$$\|\rho^\varepsilon - S\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{2} K(\rho^\varepsilon; S) \leq \left( \|S - \rho\|_{L_2(\Pi)} + \frac{\varepsilon}{2} a(\log S) \right)^2.$$

*In particular, it implies that  $\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{2} K(\rho^\varepsilon; \rho) \leq \frac{\varepsilon^2}{4} a^2(\log \rho)$ . Moreover, the following bound also holds:*

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \inf_{S \in \mathcal{S}} \left[ \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon a(\log S) \|S - \rho\|_{L_2(\Pi)} + \frac{\varepsilon^2}{2} a^2(\log S) \right].$$

**Proof.** Our starting point is the relationship (4.2) from the proof of Proposition 3. It follows from the definition of  $a(W)$ , from (4.2) and from Cauchy-Schwarz inequality that

$$2\|S - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \varepsilon K(S; \rho^\varepsilon) \leq 2\|S - \rho\|_{L_2(\Pi)} \|S - \rho^\varepsilon\|_{L_2(\Pi)} + \varepsilon a(\log S) \|S - \rho^\varepsilon\|_{L_2(\Pi)}.$$

It remains to solve the last inequality for  $\|S - \rho^\varepsilon\|_{L_2(\Pi)}$  to obtain the first bound of the proposition. The second bound is its special case with  $S = \rho$ . To prove the third bound note that, by the definition of  $\rho^\varepsilon$ , for all  $S \in \mathcal{S}$ ,

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \varepsilon \text{tr}(\rho^\varepsilon \log \rho^\varepsilon) \leq \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon \text{tr}(S \log S),$$

which implies

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon (\text{tr}(S \log S) - \text{tr}(\rho^\varepsilon \log \rho^\varepsilon)) \leq$$

$$\|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon \text{tr}(\log S (S - \rho^\varepsilon)) \leq \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon a(\log S) \|\rho^\varepsilon - S\|_{L_2(\Pi)},$$

where we used the fact that, by convexity of the function  $S \mapsto \text{tr}(S \log S)$ ,

$$\text{tr}(S \log S) - \text{tr}(\rho^\varepsilon \log \rho^\varepsilon) \leq \text{tr}(\log S (S - \rho^\varepsilon)).$$

It remains to bound  $\|\rho^\varepsilon - S\|_{L_2(\Pi)}$  from above using the first inequality of the proposition.

□

A consequence of propositions 3 and 4 is that

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq \frac{\varepsilon^2}{4} a^2(\log \rho) \wedge \varepsilon \|\log \rho\|. \quad (4.5)$$

We will now provide versions of approximation error bounds for special types of oracles  $S \in \mathcal{S}$ .

**Low Rank Oracles.** First we show how to adapt the bounds of Proposition 4 expressed in terms of the alignment coefficient  $a(\log S)$  for a full rank matrix  $S$  (for which  $\log S$  is well defined) to the case when  $S$  is an oracle of a small rank  $r < m$ . For a subspace  $L$  of  $\mathbb{C}^m$ , denote  $\Lambda(L) := \sup_{\|A\|_{L_2(\Pi)} \leq 1} \|P_L A P_L\|_2$ . Suppose that  $S \in \mathcal{S}$  is a matrix of rank  $r$ . To be specific, let  $S = \sum_{j=1}^r \gamma_j (e_j \otimes e_j)$ , where  $\gamma_j$  are positive eigenvalues of  $S$  and  $\{e_1, \dots, e_m\}$  is an orthonormal basis of  $\mathbb{C}^m$ . Let  $L$  be the linear span of the vectors  $e_1, \dots, e_r$ .

**Proposition 5** *There exists a numerical constant  $C > 0$  such that, for all  $\varepsilon > 0$ ,*

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq 2\|S - \rho\|_{L_2(\Pi)}^2 + C\varepsilon^2 \left[ \Lambda^2(L) r \log^2 \left( 1 + \frac{m}{\varepsilon \wedge 1} \right) + \mathbb{E}\|X\|^2 \right].$$

**Proof.** Note that, for all matrices  $W$  of rank  $r$  “supported” in the space  $L$  in the sense that  $W = P_L W P_L$ , we have

$$a(W) \leq \sup_{\|U\|_{L_2(\Pi)} \leq 1} \langle W, U \rangle = \sup_{\|U\|_{L_2(\Pi)} \leq 1} \langle W, P_L U P_L \rangle \leq \Lambda(L) \|W\|_2.$$

For  $\delta \in (0, 1)$ , consider  $S_\delta := (1 - \delta)S + \delta \frac{L_m}{m}$ . Then, using the fact that  $a(W + cI_m) = a(W)$ , we get

$$\log S_\delta = \sum_{j=1}^r \left( \log((1 - \delta)\gamma_j + \delta/m) - \log(\delta/m) \right) (e_j \otimes e_j) + \log(\delta/m) I_m$$

and

$$\begin{aligned} a(\log S_\delta) &= a \left( \sum_{j=1}^r \left( \log((1 - \delta)\gamma_j + \delta/m) - \log(\delta/m) \right) (e_j \otimes e_j) \right) \leq \\ &\Lambda(L) \left\| \sum_{j=1}^r \left( \log((1 - \delta)\gamma_j + \delta/m) - \log(\delta/m) \right) (e_j \otimes e_j) \right\|_2 \leq \end{aligned}$$

$$\Lambda(L) \left( \sum_{j=1}^r \log^2 \left( 1 + \frac{m\gamma_j}{\delta} \right) \right)^{1/2} \leq \Lambda(L) \sqrt{r} \log \left( 1 + \frac{m\|S\|}{\delta} \right).$$

Note also that  $\|S - S_\delta\|_{L_2(\Pi)}^2 = \delta^2 \|S - I_m/m\|_{L_2(\Pi)}^2 \leq 4\delta^2 \mathbb{E}\|X\|^2$ , since

$$\begin{aligned} \|S - I_m/m\|_{L_2(\Pi)}^2 &\leq 2(\mathbb{E}\langle S, X \rangle^2 + \mathbb{E}\langle I_m/m, X \rangle^2) \leq \\ 2(\|S\|_1^2 \mathbb{E}\|X\|^2 + \|I_m/m\|_1^2 \mathbb{E}\|X\|^2) &\leq 4\mathbb{E}\|X\|^2. \end{aligned}$$

Thus, it easily follows from Proposition 4 that

$$\begin{aligned} \|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq \frac{3}{2} \|S_\delta - \rho\|_{L_2(\Pi)}^2 + \varepsilon^2 a^2 (\log S_\delta) \leq \\ \frac{3}{2} \left( \|S - \rho\|_{L_2(\Pi)} + \|S_\delta - S\|_{L_2(\Pi)} \right)^2 &+ \Lambda^2(L) r \varepsilon^2 \log^2 \left( 1 + \frac{m}{\delta} \right) \leq \\ \frac{3}{2} \left( \frac{4}{3} \|S - \rho\|_{L_2(\Pi)}^2 + 4 \|S_\delta - S\|_{L_2(\Pi)}^2 \right) &+ \Lambda^2(L) r \varepsilon^2 \log^2 \left( 1 + \frac{m}{\delta} \right) \leq \\ 2 \|S - \rho\|_{L_2(\Pi)}^2 + 24 \mathbb{E}\|X\|^2 \delta^2 &+ \Lambda^2(L) r \varepsilon^2 \log^2 \left( 1 + \frac{m}{\delta} \right). \end{aligned}$$

Taking  $\delta = \varepsilon \wedge 1$ , this yields

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq 2 \|S - \rho\|_{L_2(\Pi)}^2 + C \varepsilon^2 \left[ \Lambda^2(L) r \log^2 \left( 1 + \frac{m}{\varepsilon \wedge 1} \right) + \mathbb{E}\|X\|^2 \right]$$

with a numerical constant  $C > 0$ .

□

**Remark.** The bound of Proposition 5 can be also written in the following form that might be preferable when  $\mathbb{E}\|X\|^2$  is large:

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq 2 \|S - \rho\|_{L_2(\Pi)}^2 + C \varepsilon^2 \left[ \Lambda^2(L) r \log^2 \left( 1 + \left( \frac{m \mathbb{E}^{1/2} \|X\|^2}{\varepsilon} \vee m \right) \right) + 1 \right].$$

In the proof, it is enough to take  $\delta := \frac{\varepsilon}{\mathbb{E}^{1/2} \|X\|^2} \wedge 1$ .

Note that if  $\{E_i, i = 1, \dots, m^2\}$  is an orthonormal basis of  $\mathbb{M}_m(\mathbb{C})$  consisting of Hermitian matrices and  $X$  is uniformly distributed in  $\{E_i, i = 1, \dots, m^2\}$ , then for all Hermitian  $A$

$$\|A\|_{L_2(\Pi)}^2 = \mathbb{E}\langle A, X \rangle^2 = m^{-2} \sum_{j=1}^{m^2} \langle A, E_j \rangle^2 = m^{-2} \|A\|_2^2.$$

Therefore  $\Lambda(L) \leq \sup_{\|A\|_{L_2(\Pi)} \leq 1} \|A\|_2 = \sup_{\|A\|_2 \leq m} \|A\|_2 = m$ . Also, in this case  $\|X\| \leq \|X\|_2 = 1$ . Thus, Proposition 5 yields

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq 2\|S - \rho\|_{L_2(\Pi)}^2 + Cm^2r\varepsilon^2 \log^2\left(1 + \frac{m}{\varepsilon \wedge 1}\right) + C\varepsilon^2.$$

**Gibbs Oracles.** Let  $H$  be a Hermitian matrix (“a Hamiltonian”) and let  $\beta > 0$ . Consider the following density matrix (a “Gibbs oracle”):

$$\rho_{H,\beta} := \frac{e^{-\beta H}}{\text{tr}(e^{-\beta H})}.$$

For simplicity, assume in what follows that  $\beta = 1$  (in fact, one can always replace  $H$  by  $\beta H$ ) and denote  $\rho_H := \frac{e^{-H}}{\text{tr}(e^{-H})}$ . Let  $\gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_m$  be the eigenvalues of  $H$  and  $e_1, \dots, e_m$  be the corresponding eigenvectors. Let  $L_r = \text{l.s.}(\{e_1, \dots, e_r\})$  and  $H_{\leq r} := \sum_{j=1}^r \gamma_j(e_j \otimes e_j)$ ,  $H_{>r} := \sum_{j=r+1}^m \gamma_j(e_j \otimes e_j)$ . It is easy to see that

$$\|P_{L_r^\perp} \rho_H P_{L_r^\perp}\|_1 = \frac{\sum_{k \geq r+1} e^{-\gamma_k}}{\sum_{k \geq 1} e^{-\gamma_k}} =: \delta_r(H).$$

Under reasonable conditions on the spectrum of  $H$ , the quantity  $\delta_r(H)$  decreases fast enough when  $r$  increases. Thus,  $\rho_H$  can be well approximated by low rank matrices.

The next statement follows immediately from Proposition 4. Here the unknown density matrix  $\rho$  is approximated by a Gibbs model with an arbitrary Hamiltonian. The error is controlled in terms of the  $L_2(\Pi)$ -distance between  $\rho$  and the oracle  $\rho_H$  and also in terms of the alignment coefficient  $a(H_{\leq r})$  for a “low rank part”  $H_{\leq r}$  of the Hamiltonian  $H$  and the quantity  $\delta_r(H)$ .

**Proposition 6** *For all Hermitian nonnegatively definite matrices  $H$  and for all  $\varepsilon > 0$ ,*

$$\|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq 2\|\rho_H - \rho\|_{L_2(\Pi)}^2 + 24 \max_{1 \leq k \leq m} \mathbb{E}\langle X e_k, e_k \rangle^2 \delta_r^2(H) + a^2(H_{\leq r}) \varepsilon^2.$$

**Proof.** We will use the last bound of proposition 4 with  $S = \rho_{H_{\leq r}}$ . Note that

$$a(\log \rho_{H_{\leq r}}) = a(-H_{\leq r} - \log \text{tr}(e^{-H_{\leq r}}) I_m) = a(H_{\leq r}).$$

Therefore, we have

$$\begin{aligned} \|\rho^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq \|\rho_{H_{\leq r}} - \rho\|_{L_2(\Pi)}^2 + \varepsilon a(H_{\leq r}) \|\rho_{H_{\leq r}} - \rho\|_{L_2(\Pi)} + \frac{\varepsilon^2}{2} a^2(H_{\leq r}) \leq \\ &\frac{3}{2} \|\rho_{H_{\leq r}} - \rho\|_{L_2(\Pi)}^2 + \varepsilon^2 a^2(H_{\leq r}). \end{aligned}$$



In addition to this,

$$\|\rho_H - \rho_{H_{\leq r}}\|_{L_2(\Pi)} = \left\| \frac{\sum_{k=1}^m e^{-\gamma_k} (e_k \otimes e_k)}{\sum_{k=1}^m e^{-\gamma_k}} - \frac{\sum_{k=1}^r e^{-\gamma_k} (e_k \otimes e_k)}{\sum_{k=1}^r e^{-\gamma_k}} \right\|_{L_2(\Pi)},$$

which can be easily bounded from above by

$$2\delta_r(H) \max_{1 \leq k \leq m} \|e_k \otimes e_k\|_{L_2(\Pi)} = 2\delta_r(H) \max_{1 \leq k \leq m} \mathbb{E}^{1/2} \langle X e_k, e_k \rangle^2.$$

The result follows immediately (by the same argument as in the proof of Proposition 5).  $\square$

## 5 Random Error Bounds and Oracle Inequalities

We now turn to the analysis of random error of the estimator  $\hat{\rho}^\varepsilon$ . We obtain upper bounds on the  $L_2(\Pi)$  and Kullback-Leibler distances of this estimator to an arbitrary oracle  $S \in \mathcal{S}$  of full rank. In particular, this includes bounding the distances between  $\hat{\rho}^\varepsilon$  and  $\rho^\varepsilon$ . As a consequence, we will obtain **oracle inequalities** for the empirical solution  $\hat{\rho}^\varepsilon$ . The size of both errors  $\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2$  and  $K(\hat{\rho}^\varepsilon; S)$  will be controlled in terms of the squared  $L_2(\Pi)$ -distance  $\|S - \rho\|_{L_2(\Pi)}^2$  from the oracle to the target density matrix  $\rho$  and also in terms of such characteristics of the oracle as the norm  $\|\log S\|$  or the alignment coefficient  $a(\log S)$  that have been already used in the approximation error bounds of the previous section (see propositions 3, 4). However, in the case of the random error, we also need some additional quantities that describe the properties of the design distribution  $\Pi$  and of the noise  $\xi$ . These quantities are explicitly involved in the statements of the results below which makes these statements somewhat complicated. At the same time, it is easy to control these quantities in concrete examples and to derive in special cases the bounds that are easier to understand.

**Assumptions on the design distribution  $\Pi$ .** In this section, it will be assumed that  $X$  is a random Hermitian  $m \times m$  matrix and that, for some constant  $U > 0$ ,  $\|X\| \leq U$ . We will denote

$$\sigma_X^2 := \|\mathbb{E}(X - \mathbb{E}X)^2\|, \quad \sigma_{X \otimes X}^2 := \|\mathbb{E}(X \otimes X - \mathbb{E}(X \otimes X))^2\|.$$

Let  $L \subset \mathbb{C}^m$  be a subspace of dimension  $r \leq m$  and let  $\mathcal{P}_L : \mathbb{M}_m(\mathbb{C}) \mapsto \mathbb{M}_m(\mathbb{C})$ ,  $\mathcal{P}_L x := x - P_{L^\perp} x P_{L^\perp}$ . We will use the following quantity:

$$\beta(L) := \sup_{\|A\|_{L_2(\Pi)} \leq 1} \|\mathcal{P}_L A\|_{L_2(\Pi)}.$$

Note that  $\|\mathcal{P}_L A\|_2 \leq \|A\|_2$  (for a proof, choose a basis  $\{e_1, \dots, e_m\}$  of  $\mathbb{C}^m$  such that  $L = \text{l.s.}(e_1, \dots, e_r)$  and represent the linear transformations in this basis). If, for all  $A$ ,  $K_1\|A\|_2 \leq \|A\|_{L_2(\Pi)} \leq K_2\|A\|_2$ , then  $\beta(L) \leq K_2/K_1$ . In particular, if  $K_1 = K_2$ , then  $\beta(L) = 1$  (which is the case, for instance, when  $X$  is sampled at random from an orthonormal basis).

**Assumptions on the noise  $\xi$ .** Recall that  $\mathbb{E}\xi = 0$  and let  $\sigma_\xi^2 := \mathbb{E}\xi^2 < +\infty$ . We will further assume that the noise is uniformly bounded by a constant  $c_\xi > 0 : |\xi| \leq c_\xi$ , and the proofs of the results of this section will be given under this assumption. Alternatively, one can assume that the noise is not necessarily uniformly bounded, but  $\|\xi\|_{\psi_1} < +\infty$ . This includes, for instance, the case of Gaussian noise. For such an unbounded noise, one should replace in the proofs of theorems 4, 5 and 6 below the noncommutative Bernstein inequality of Ahlswede and Winter by the bound of Proposition 2. One should also use a version of concentration inequality for empirical processes by Adamczak (2008) instead of the usual version of Talagrand for bounded function classes (see Section 3).

Given  $t > 0$ , denote  $t_m := t + \log(2m)$ ,  $\tau_n := t + \log \log_2(2n)$  and

$$\varepsilon_{n,m} := (\sigma_\xi \sigma_X \vee \sigma_\xi \|\mathbb{E}X\| \vee \sigma_{X \otimes X}) \sqrt{\frac{t_m}{n}} \bigvee (c_\xi U \vee U^2) \frac{t_m}{n}.$$

We will start with a simple result in spirit of approximation error bound of Proposition 3.

**Theorem 4** *There exists a constant  $C > 0$  such that, for all  $S \in \mathcal{S}$  and for all  $\varepsilon \geq 0$ , with probability at least  $1 - e^{-t}$*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 &\leq \|S - \rho\|_{L_2(\Pi)}^2 + C \left[ \varepsilon (\|\log S\| \wedge \log \Gamma) \bigvee \|S - \rho\|_{L_2(\Pi)} U \sqrt{\frac{t_m}{n}} \bigvee \right. \\ &\quad \left. (\sigma_\xi \sigma_X \vee \sigma_\xi \|\mathbb{E}X\| \vee \sigma_{X \otimes X}) \sqrt{\frac{t_m}{n}} \bigvee (c_\xi U \vee U^2) \frac{t_m}{n} \right] \end{aligned} \quad (5.1)$$

and

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq \|S - \rho\|_{L_2(\Pi)}^2 + C \left[ \varepsilon (\|\log S\| \wedge \log \Gamma) \bigvee \|S - \rho\|_{L_2(\Pi)} U \sqrt{\frac{t_m}{n}} \bigvee \right. \\ &\quad \left. (\sigma_\xi \sigma_X \vee \sigma_\xi \|\mathbb{E}X\| \vee \sigma_{X \otimes X}) \sqrt{\frac{t_m}{n}} \bigvee (c_\xi U \vee U^2) \frac{t_m}{n} \right], \end{aligned} \quad (5.2)$$

where  $\Gamma := \frac{m\mathbb{E}^{1/2}\|X\|^2}{\sqrt{\varepsilon}} \vee m$ . In particular,

$$\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq C \left[ \varepsilon (\|\log \rho\| \wedge \log \Gamma) \bigvee (\sigma_\xi \sigma_X \vee \sigma_\xi \|\mathbb{E}X\| \vee \sigma_{X \otimes X}) \sqrt{\frac{t_m}{n}} \bigvee (c_\xi U \vee U^2) \frac{t_m}{n} \right]. \quad (5.3)$$

Note that this result holds for all  $\varepsilon \geq 0$ , including the case of  $\varepsilon = 0$  that corresponds to the least squares estimator over the set  $\mathcal{S}$  of all density matrices. The approximation error term  $\|\log S\|_\varepsilon$  in the bounds of Theorem 4 is of the order  $O(\varepsilon)$  (as in Proposition 3) and the random error terms are, up to logarithmic factors, of the order  $O(\frac{1}{\sqrt{n}})$  with respect to the sample size  $n$ .

The next result provides a more subtle oracle inequality that is akin to approximation error bounds of Proposition 4. In this oracle inequality, the approximation error term due to von Neumann entropy penalization is  $a^2(\log S)\varepsilon^2$  (as in Proposition 4), so, it is of the order  $O(\varepsilon^2)$ . Note that it is assumed implicitly that  $a^2(\log S) < +\infty$ , i.e., that  $S$  is of full rank and the matrix  $\log S$  is well defined. The random error terms are of the order  $O(n^{-1})$  as  $n \rightarrow \infty$  (up to logarithmic factors) with an exception of the term  $\sigma_\xi(\sigma_X \vee \|\mathbb{E}X\|)\|P_{L^\perp}SP_{L^\perp}\|_1\sqrt{\frac{t_m}{n}}$ , which depends on how well the oracle  $S$  is approximated by low rank matrices. If  $\|P_{L^\perp}SP_{L^\perp}\|_1$  is small, say of the order  $n^{-1/2}$  for a subspace  $L$  of a small dimension  $r$ , this term becomes comparable to other terms in the bound, or even smaller. The inequalities hold only for the values of regularization parameter  $\varepsilon$  above certain threshold (so, this result does not apply to the simple least squares estimator). The first bound shows that if there is an oracle  $S \in \mathcal{S}$  such that: (a) it is “well aligned”, that is,  $a(\log S)$  is small; (b) there exists a subspace  $L$  of small dimension  $r$  such that the oracle matrix  $S$  is “almost supported” in  $L$ , that is,  $\|P_{L^\perp}SP_{L^\perp}\|_1$  is small; and (c)  $S$  provides a good approximation of the density matrix  $\rho$ , that is,  $\|S - \rho\|_{L_2(\Pi)}^2$  is small, then the empirical solution  $\hat{\rho}^\varepsilon$  will be in the intersection of the  $L_2(\Pi)$ -ball and the Kullback-Leibler “ball” of small enough radii around the oracle  $S$ . The second bound is an oracle inequality showing how the  $L_2(\Pi)$ -error  $\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2$  depends on the properties of the oracle  $S$ .

**Theorem 5** *There exist numerical constants  $C > 0, D > 0$  such that the following holds. For all  $t > 0$ , for all  $\lambda > 0$ , for all  $\varepsilon \geq D\varepsilon_{n,m}$ , for all subspaces  $L \subset \mathbb{C}^m$  with  $\dim(L) := r$ , and for all  $S \in \mathcal{S}$ , with probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{4}K(\hat{\rho}^\varepsilon; S) &\leq (1 + \lambda)\|S - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ a^2(\log S)\varepsilon^2 \vee \right. \\ &\left. \sigma_\xi^2\beta^2(L)\frac{mr + \tau_n}{n} \vee \sigma_\xi(\sigma_X \vee \|\mathbb{E}X\|)\|P_{L^\perp}SP_{L^\perp}\|_1\sqrt{\frac{t_m}{n}} \vee c_\xi U \frac{\tau_n \vee t_m}{n} \vee U^2 \frac{t_m}{n} \right] \end{aligned} \quad (5.4)$$

and

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq (1 + \lambda)\|S - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ a^2(\log S)\varepsilon^2 \bigvee \sigma_\xi^2 \beta^2(L) \frac{mr + \tau_n}{n} \bigvee \right. \\ &\quad \left. \sigma_\xi(\sigma_X \vee \|\mathbb{E}X\|) \|P_{L^\perp} S P_{L^\perp}\|_1 \sqrt{\frac{t_m}{n}} \bigvee c_\xi U \frac{\tau_n \vee t_m}{n} \bigvee U^2 \frac{t_m}{n} \right]. \end{aligned} \quad (5.5)$$

Next we give a version of (5.4) in a special case when  $S = \rho^\varepsilon$ . This provides bounds on random errors of estimation of the true penalized solution  $\rho^\varepsilon$  by its empirical version  $\hat{\rho}^\varepsilon$  both in the  $L_2(\Pi)$  and in the Kullback-Leibler distances. Note that unlike the bounds for an arbitrary oracle  $S$ , there is no dependence on the alignment coefficient  $a(\log \rho^\varepsilon)$  in this case. The result essentially shows that as soon as the true solution  $\rho^\varepsilon$  is approximately low rank in the sense that  $P_{L^\perp} \rho^\varepsilon P_{L^\perp}$  is “small” for a subspace  $L$  of a “small” dimension  $r$  and  $\rho^\varepsilon$  provides a good approximation of the target density matrix  $\rho$ , the empirical solution  $\hat{\rho}^\varepsilon$  would also provide a good approximation of  $\rho$  and it would be approximately low rank.

**Theorem 6** *There exist numerical constants  $C > 0, D > 0$  such that the following holds. For all  $t > 0$ , for all  $\varepsilon \geq D\varepsilon_{n,m}$  and for all subspaces  $L \subset \mathbb{C}^m$  with  $\dim(L) := r$ , with probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned} &\|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; \rho^\varepsilon) \leq \\ &C \left[ \sigma_\xi^2 \beta^2(L) \frac{mr + \tau_n}{n} \bigvee \sigma_\xi(\sigma_X \vee \|\mathbb{E}X\|) \|P_{L^\perp} \rho^\varepsilon P_{L^\perp}\|_1 \sqrt{\frac{t_m}{n}} \bigvee \right. \\ &\quad \left. U \|\rho^\varepsilon - \rho\|_{L_2(\Pi)} \sqrt{\frac{t_m}{n}} \bigvee U^2 \|\rho^\varepsilon - \rho\|_1 \frac{t_m}{n} \bigvee c_\xi U \frac{\tau_n \vee t_m}{n} \right]. \end{aligned} \quad (5.6)$$

**Remark.** In the case when the noise is not necessarily bounded, but  $\|\xi\|_{\psi_1} < +\infty$ , the results still hold with the following simple modifications. In bounds (5.1), (5.2), (5.3) and in the definition of  $\varepsilon_{n,m}$ , the term  $(c_\xi U \vee U^2) \frac{t_m}{n}$  is to be replaced by

$$\left( \|\xi\|_{\psi_1} U \log \left( \frac{\|\xi\|_{\psi_1}}{\sigma_\xi} \frac{U}{\sigma_X} \right) \bigvee U^2 \right) \frac{t_m}{n}.$$

In the bounds of theorems 5 and 6, the term  $c_\xi U \frac{\tau_n \vee t_m}{n}$  is to be replaced by

$$\|\xi\|_{\psi_1} U \frac{\tau_n \log n}{n} \bigvee \|\xi\|_{\psi_1} U \log \left( \frac{\|\xi\|_{\psi_1}}{\sigma_\xi} \frac{U}{\sigma_X} \right) \frac{t_m}{n}.$$

We will provide a detailed proof of Theorem 5. The proof of Theorem 4 is its simplified version. The proof of Theorem 6 relies on the bounds derived in the proof

of Theorem 5. It is also possible to derive the oracle inequalities of Theorem 5 from Theorem 6 and from the approximation error bounds of Proposition 4. Throughout the proofs below,  $C, C_1, \dots$  are numerical constants whose values might be different in different places.

**Proof of Theorem 5.** Denote

$$L_n(S) := n^{-1} \sum_{j=1}^n (Y_j - \text{tr}(SX_j))^2 + \varepsilon \text{tr}(S \log S).$$

For any  $S \in \mathcal{S}$  of full rank and any direction  $\nu \in \mathbb{M}_m(\mathbb{C})$ , we have

$$DL_n(S; \nu) = 2n^{-1} \sum_{j=1}^n (\langle S, X_j \rangle - Y_j) \langle \nu, X_j \rangle + \varepsilon \text{tr}(\nu \log S).$$

By necessary conditions of extrema in the convex optimization problem (1.2),  $DL_n(\hat{\rho}^\varepsilon; \hat{\rho}^\varepsilon - S) \leq 0$ , which implies

$$DL(\hat{\rho}^\varepsilon; \hat{\rho}^\varepsilon - S) - DL(S; \hat{\rho}^\varepsilon - S) \leq -DL(S; \hat{\rho}^\varepsilon - S) + DL(\hat{\rho}^\varepsilon; \hat{\rho}^\varepsilon - S) - DL_n(\hat{\rho}^\varepsilon; \hat{\rho}^\varepsilon - S). \quad (5.7)$$

Note that

$$DL(\hat{\rho}^\varepsilon; \hat{\rho}^\varepsilon - S) - DL(S; \hat{\rho}^\varepsilon - S) = 2\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; S)$$

(see the proof of Proposition 3) and

$$DL(S; \hat{\rho}^\varepsilon - S) = 2\langle S - \rho, \hat{\rho}^\varepsilon - S \rangle_{L_2(\Pi)} + \varepsilon \text{tr}((\hat{\rho}^\varepsilon - S) \log S).$$

By a simple algebra similar to what has been already used in the proofs of propositions 3, 4, we get the following bound:

$$\begin{aligned} & 2\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + 2\langle S - \rho, \hat{\rho}^\varepsilon - S \rangle_{L_2(\Pi)} + \varepsilon K(\hat{\rho}^\varepsilon; S) = \\ & \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 - \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; S) \leq \\ & -\varepsilon \text{tr}((\hat{\rho}^\varepsilon - S) \log S) - \frac{2}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right) + \\ & \frac{2}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) - \frac{2}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, X_j \rangle. \end{aligned} \quad (5.8)$$

Since  $\varepsilon |\text{tr}((\hat{\rho}^\varepsilon - S) \log S)| \leq \varepsilon a(\log S) \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}$ , we get from (5.8) that

$$\begin{aligned} & \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; S) \leq \\ & \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon a(\log S) \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} - \frac{2}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right) + \\ & \frac{2}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) - \frac{2}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, X_j \rangle. \end{aligned} \quad (5.9)$$

We need to bound the empirical processes in the right hand side of bound (5.9). We will do it in several steps by bounding each term separately.

**Step 1.** To bound the first term note that

$$\frac{1}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right) = \left\langle (\hat{\rho}^\varepsilon - S) \otimes (\hat{\rho}^\varepsilon - S), \frac{1}{n} \sum_{j=1}^n ((X_j \otimes X_j) - \mathbb{E}(X \otimes X)) \right\rangle.$$

Therefore,

$$\left| \frac{1}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right) \right| \leq \|\hat{\rho}^\varepsilon - S\|_1^2 \left\| \frac{1}{n} \sum_{j=1}^n ((X_j \otimes X_j) - \mathbb{E}(X \otimes X)) \right\|.$$

Note that  $\|X \otimes X\| = \|X\|^2 \leq U^2$  and also  $\|X \otimes X - \mathbb{E}(X \otimes X)\| \leq 2U^2$ . Using noncommutative Bernstein's inequality (see (3.2) in subsection 3.3) we can claim that with probability at least  $1 - e^{-t}$

$$\left\| \frac{1}{n} \sum_{j=1}^n ((X_j \otimes X_j) - \mathbb{E}(X \otimes X)) \right\| \leq 4 \left( \sigma_{X \otimes X} \sqrt{\frac{t + \log(2m^2)}{n}} \vee U^2 \frac{t + \log(2m^2)}{n} \right)$$

and, with the same probability,

$$\begin{aligned} & \left| \frac{1}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right) \right| \leq \\ & 4 \left( \sigma_{X \otimes X} \sqrt{\frac{t + \log(2m^2)}{n}} \vee U^2 \frac{t + \log(2m^2)}{n} \right) \|\hat{\rho}^\varepsilon - S\|_1^2. \end{aligned}$$

**Step 2.** The second term can be written as

$$\frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) =$$

$$\left\langle \hat{\rho}^\varepsilon - S, \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle X_j - \mathbb{E} \langle S - \rho, X \rangle X \right) \right\rangle$$

and bounded as follows

$$\left| \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) \right| \leq$$

$$\| \hat{\rho}^\varepsilon - S \|_1 \left\| \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle X_j - \mathbb{E} \langle S - \rho, X \rangle X \right) \right\|.$$

We use again the noncommutative version of Bernstein's inequality to show that with probability at least  $1 - e^{-t}$

$$\left\| \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle X_j - \mathbb{E} \langle S - \rho, X \rangle X \right) \right\| \leq$$

$$4U \|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{t + \log(2m)}{n}} \vee 4U^2 \|S - \rho\|_1 \frac{t + \log(2m)}{n},$$

where we also used simple bounds  $\|\mathbb{E} \langle S - \rho, X \rangle^2 X^2\| \leq U^2 \|S - \rho\|_{L_2(\Pi)}^2$  and  $\|\langle S - \rho, X \rangle X\| \leq U^2 \|S - \rho\|_1$ . Since  $\|\hat{\rho}^\varepsilon - S\|_1 \leq 2$ , we get

$$\left| \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) \right| \leq$$

$$8U \|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{t + \log(2m)}{n}} \vee 8U^2 \|S - \rho\|_1 \frac{t + \log(2m)}{n}.$$

**Step 3.** We turn now to bounding the third term in the right hand side of (5.9). It is easy to decompose it as follows:

$$\frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, X_j \rangle = \left\langle P_{L^\perp} (\hat{\rho}^\varepsilon - S) P_{L^\perp}, \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\rangle +$$

$$\frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, \mathcal{P}_L X_j \rangle. \quad (5.10)$$

Note that

$$\left| \left\langle P_{L^\perp} (\hat{\rho}^\varepsilon - S) P_{L^\perp}, \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\rangle \right| \leq \|P_{L^\perp} (\hat{\rho}^\varepsilon - S) P_{L^\perp}\|_1 \left\| \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\|.$$

Applying the noncommutative version of Bernstein's inequality one more time, we have that with probability at least  $1 - e^{-t}$

$$\left\| \frac{1}{n} \sum_{j=1}^n \xi_j (P_{L^\perp} X_j P_{L^\perp} - \mathbb{E} P_{L^\perp} X P_{L^\perp}) \right\| \leq 2\sigma_\xi \sigma_X \sqrt{\frac{t + \log(2m)}{n}} \vee 2c_\xi U \frac{t + \log(2m)}{n},$$

where we used a simple bound  $\|\mathbb{E}(P_{L^\perp}(X - \mathbb{E}X)P_{L^\perp})^2\| \leq \|\mathbb{E}(X - \mathbb{E}X)^2\| = \sigma_X^2$ . Also, it follows from the classical Bernstein's inequality and the bound  $\|\mathbb{E}(P_{L^\perp} X P_{L^\perp})\| \leq \|\mathbb{E}X\|$  that with probability at least  $1 - e^{-t}$

$$\left\| \frac{1}{n} \sum_{j=1}^n \xi_j \mathbb{E} P_{L^\perp} X P_{L^\perp} \right\| = \left| \frac{1}{n} \sum_{j=1}^n \xi_j \right| \left\| \mathbb{E} P_{L^\perp} X P_{L^\perp} \right\| \leq 2\sigma_\xi \|\mathbb{E}X\| \sqrt{\frac{t}{n}} \vee 2c_\xi \|\mathbb{E}X\| \frac{t}{n}.$$

Hence, with probability at least  $1 - 2e^{-t}$ ,

$$\begin{aligned} & \left| \left\langle P_{L^\perp}(\hat{\rho}^\varepsilon - S)P_{L^\perp}, \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\rangle \right| \leq \\ & 2\|P_{L^\perp}(\hat{\rho}^\varepsilon - S)P_{L^\perp}\|_1 \left[ \sigma_\xi(\sigma_X + \|\mathbb{E}X\|) \sqrt{\frac{t + \log(2m)}{n}} \vee 2c_\xi U \frac{t + \log(2m)}{n} \right]. \end{aligned}$$

To bound the second term in the right hand side of (5.10), denote

$$\alpha_n(\delta) := \sup_{\rho_1, \rho_2 \in \mathcal{S}, \|\rho_1 - \rho_2\|_{L_2(\Pi)} \leq \delta} \left| \frac{1}{n} \sum_{j=1}^n \xi_j \langle \rho_1 - \rho_2, \mathcal{P}_L X_j \rangle \right|.$$

Clearly,  $\left| \frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, \mathcal{P}_L X_j \rangle \right| \leq \alpha_n(\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)})$ . To control  $\alpha_n(\delta)$ , we use Talagrand's concentration inequality for empirical processes. It implies that, for all  $\delta > 0$ , with probability at least  $1 - e^{-s}$ ,

$$\alpha_n(\delta) \leq 2 \left[ \mathbb{E} \alpha_n(\delta) + \sigma_\xi \beta(L) \delta \sqrt{\frac{s}{n}} + 4c_\xi U \frac{s}{n} \right]. \quad (5.11)$$

Here we used the facts that  $\mathbb{E} \xi^2 \langle \rho_1 - \rho_2, \mathcal{P}_L X \rangle^2 \leq \sigma_\xi^2 \beta^2(L) \|\rho_1 - \rho_2\|_{L_2(\Pi)}^2$  and

$$\left| \xi \langle \rho_1 - \rho_2, \mathcal{P}_L X \rangle \right| \leq c_\xi \|\rho_1 - \rho_2\|_1 \|\mathcal{P}_L X\| \leq 2c_\xi (\|X\| + \|P_{L^\perp} X P_{L^\perp}\|) \leq 4c_\xi \|X\| \leq 4c_\xi U.$$

We will make the bound on  $\alpha_n(\delta)$  uniform in  $\delta \in [Un^{-1}, 2U]$ . To this end, we apply bound (5.11) for  $\delta = \delta_j = 2^{-j+1}U$ ,  $j = 0, 1, \dots$  and with  $s = \tau_n := t + \log \log_2(2n)$ . The union bound and the monotonicity of  $\alpha_n(\delta)$  with respect to  $\delta$  implies that with probability at least  $1 - e^{-t}$  for all  $\delta \in [Un^{-1}, 2U]$

$$\alpha_n(\delta) \leq C \left[ \mathbb{E} \alpha_n(\delta) + \sigma_\xi \beta(L) \delta \sqrt{\frac{\tau_n}{n}} + c_\xi U \frac{\tau_n}{n} \right], \quad (5.12)$$



where  $C > 0$  is a numerical constant. Now it remains to bound the expected value  $\mathbb{E}\alpha_n(\delta)$ . Let  $e_1, \dots, e_m$  be the orthonormal basis of  $\mathbb{C}^m$  such that  $L = \text{l.s.}\{e_1, \dots, e_r\}$ . Denote  $E_{ij}(x)$  the entries of the linear transformation  $x \in \mathbb{M}_m(\mathbb{C})$  in this basis. Clearly, the function  $\langle \rho_1 - \rho_2, \mathcal{P}_L x \rangle$  belongs to the space  $\mathcal{L} := \text{l.s.}\{E_{ij} : i \leq r \text{ or } j \leq r\}$  of dimension  $m^2 - (m-r)^2 = 2mr - r^2$ . Therefore,

$$\mathbb{E}\alpha_n(\delta) \leq \mathbb{E} \sup_{f \in \mathcal{L}, \|f\|_{L_2(\Pi)} \leq \beta(L)\delta} \left| \frac{2}{n} \sum_{j=1}^n \xi_j f(X_j) \right|.$$

Using standard bounds for empirical processes indexed by finite dimensional function classes, we get  $\mathbb{E}\alpha_n(\delta) \leq 2\sqrt{2}\sigma_\xi\beta(L)\delta\sqrt{\frac{mr}{n}}$ . We can conclude that the following bound on  $\alpha_n(\delta)$  holds with probability at least  $1 - e^{-t}$  for all  $\delta \in [Un^{-1}, 2U]$  :

$$\alpha_n(\delta) \leq C \left[ \sigma_\xi\beta(L)\delta\sqrt{\frac{mr}{n}} + \sigma_\xi\beta(L)\delta\sqrt{\frac{\tau_n}{n}} + c_\xi U \frac{\tau_n}{n} \right]. \quad (5.13)$$

Note that since  $\|\hat{\rho}^\varepsilon - S\|_1 \leq 2$  and  $\|X\| \leq U$ , we have  $\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 = \mathbb{E}\langle \hat{\rho}^\varepsilon - S, X \rangle \leq 4U^2$ , so,  $\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} \leq 2U$ . As a result, with probability at least  $1 - e^{-t}$ , we either have  $\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} < Un^{-1}$ , or

$$\left| \frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, \mathcal{P}_L X_j \rangle \right| \leq C \left[ \sigma_\xi\beta(L)\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}\sqrt{\frac{mr}{n}} + \sigma_\xi\beta(L)\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}\sqrt{\frac{\tau_n}{n}} + c_\xi U \frac{\tau_n}{n} \right].$$

In the first case, we still have

$$\left| \frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, \mathcal{P}_L X_j \rangle \right| \leq C \left[ \sigma_\xi\beta(L)\frac{U}{n}\sqrt{\frac{mr}{n}} + \sigma_\xi\beta(L)\frac{U}{n}\sqrt{\frac{\tau_n}{n}} + c_\xi U \frac{\tau_n}{n} \right].$$

Let us assume in what follows that  $\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} \geq Un^{-1}$  since another case is even easier to handle.

We now substitute the bounds of steps 1–3 in the right hand side of (5.9) to get the following inequality that holds with some constant  $C > 0$  and with probability at least  $1 - 5e^{-t}$  :

$$\begin{aligned} & \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; S) \leq \\ & \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon a(\log S)\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} + \\ & 16 \left( \sigma_{X \otimes X} \sqrt{\frac{t_m}{n}} \vee U^2 \frac{t_m}{n} \right) \|\hat{\rho}^\varepsilon - S\|_1^2 + 16U\|S - \rho\|_{L_2(\Pi)}\sqrt{\frac{t_m}{n}} \vee 16U^2 \frac{t_m}{n} + \\ & 4\|P_{L^\perp}(\hat{\rho}^\varepsilon - S)P_{L^\perp}\|_1 \left[ \sigma_\xi(\sigma_X + \|\mathbb{E}X\|)\sqrt{\frac{t_m}{n}} \vee 2c_\xi U \frac{t_m}{n} \right] + \\ & C \left[ \sigma_\xi\beta(L)\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}\sqrt{\frac{mr + \tau_n}{n}} \vee c_\xi U \frac{\tau_n}{n} \right]. \end{aligned} \quad (5.14)$$

Under the assumption  $\varepsilon \geq D\varepsilon_{n,m}$  with a sufficiently large constant  $D > 0$ , it is easy to get that

$$16 \left( \sigma_{X \otimes X} \sqrt{\frac{t_m}{n}} \bigvee U^2 \frac{t_m}{n} \right) \|\hat{\rho}^\varepsilon - S\|_1^2 \leq \frac{\varepsilon}{2} \|\hat{\rho}^\varepsilon - S\|_1^2 \leq \frac{\varepsilon}{2} K(\hat{\rho}^\varepsilon; S). \quad (5.15)$$

Also, by Proposition 1,

$$\|P_{L^\perp}(\hat{\rho}^\varepsilon - S)P_{L^\perp}\|_1 \leq \|P_{L^\perp}\hat{\rho}^\varepsilon P_{L^\perp}\|_1 + \|P_{L^\perp}SP_{L^\perp}\|_1 \leq 3\|P_{L^\perp}SP_{L^\perp}\|_1 + 2K(\hat{\rho}^\varepsilon; S),$$

and, under the same assumption that  $\varepsilon \geq D\varepsilon_{n,m}$  with a sufficiently large constant  $D > 0$ ,

$$\begin{aligned} 4\|P_{L^\perp}(\hat{\rho}^\varepsilon - S)P_{L^\perp}\|_1 \left[ \sigma_\xi(\sigma_X + \|\mathbb{E}X\|) \sqrt{\frac{t_m}{n}} \bigvee 2c_\xi U \frac{t_m}{n} \right] &\leq \\ C\|P_{L^\perp}SP_{L^\perp}\|_1 \left[ \sigma_\xi(\sigma_X \vee \|\mathbb{E}X\|) \sqrt{\frac{t_m}{n}} \bigvee c_\xi U \frac{t_m}{n} \right] &+ \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S). \end{aligned} \quad (5.16)$$

Combining bounds (5.15) and (5.16) with (5.14) yields

$$\begin{aligned} \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S) &\leq \\ \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon a(\log S) \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} + \\ C \left[ \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} \sigma_\xi \beta(L) \sqrt{\frac{mr + \tau_n}{n}} \bigvee U \|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{t_m}{n}} \bigvee \right. \\ \left. \|P_{L^\perp}SP_{L^\perp}\|_1 \sigma_\xi(\sigma_X \vee \|\mathbb{E}X\|) \sqrt{\frac{t_m}{n}} \bigvee c_\xi U \frac{\tau_n \vee t_m}{n} \bigvee U^2 \frac{t_m}{n} \right] \end{aligned} \quad (5.17)$$

with some constant  $C > 0$ . It follows from the last inequality that

$$\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 \leq A \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} + B - \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S), \quad (5.18)$$

where  $A := \frac{\varepsilon}{2} a(\log S) + C \sigma_\xi \beta(L) \sqrt{\frac{mr + \tau_n}{n}}$  and

$$\begin{aligned} B := \|S - \rho\|_{L_2(\Pi)}^2 - \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \\ C \left[ \|S - \rho\|_{L_2(\Pi)} U \sqrt{\frac{t_m}{n}} \bigvee \|P_{L^\perp}SP_{L^\perp}\|_1 \sigma_\xi(\sigma_X \vee \|\mathbb{E}X\|) \sqrt{\frac{t_m}{n}} \bigvee c_\xi U \frac{\tau_n \vee t_m}{n} \bigvee U^2 \frac{t_m}{n} \right]. \end{aligned}$$

It is easy to check that

$$\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 \leq \left( \frac{A + \sqrt{A^2 + 4(B - (\varepsilon/4)K(\hat{\rho}^\varepsilon; S))}}{2} \right)^2 \leq \left( A + \sqrt{\left( B - \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S) \right)_+} \right)^2.$$

If  $\frac{\varepsilon}{4}K(\hat{\rho}^\varepsilon; S) \geq B$ , then  $\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 \leq A^2$ , which, in view of (5.18), implies

$$\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{4}K(\hat{\rho}^\varepsilon; S) \leq A^2 + B.$$

Otherwise, we have  $\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 \leq A^2 + 2A\sqrt{B} + B - \frac{\varepsilon}{4}K(\hat{\rho}^\varepsilon; S)$ , which, for all  $\lambda > 0$ , implies

$$\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{4}K(\hat{\rho}^\varepsilon; S) \leq \left(\frac{2}{\lambda} + 1\right)A^2 + (1 + \lambda/2)B.$$

In both cases, by the definitions of  $A$  and  $B$  and by elementary algebra, one can easily get the bound

$$\begin{aligned} & \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{4}K(\hat{\rho}^\varepsilon; S) \leq \\ & (1 + \lambda)\|S - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ a^2(\log S)\varepsilon^2 \bigvee \sigma_\xi^2 \beta^2(L) \frac{mr + \tau_n}{n} \bigvee \right. \\ & \left. \sigma_\xi(\sigma_X \vee \|\mathbb{E}X\|) \|P_{L^\perp} S P_{L^\perp}\|_1 \sqrt{\frac{t_m}{n}} \bigvee c_\xi U \frac{\tau_n \vee t_m}{n} \bigvee U^2 \frac{t_m}{n} \right] \end{aligned} \quad (5.19)$$

that holds with probability at least  $1 - 5e^{-t}$  and with a sufficiently large constant  $C$ . To replace the probability  $1 - 5e^{-t}$  by  $1 - e^{-t}$ , it is enough to replace  $t$  by  $t + \log 5$  and to adjust the values of constants  $C, D$  accordingly.

□

**Proof of Theorem 4.** We get back to bound (5.8) in the proof of Theorem 5. This time, we bound the term  $\text{tr}((\hat{\rho}^\varepsilon - S) \log S)$  in (5.8) in a slightly different way

$$|\text{tr}((\hat{\rho}^\varepsilon - S) \log S)| \leq \|\log S\| \|\hat{\rho}^\varepsilon - S\|_1 \leq 2\|\log S\|,$$

which leads to the following bound (instead of bound (5.9)):

$$\begin{aligned} & \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; S) \leq \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon \|\log S\| + \\ & - \frac{1}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right) + \\ & \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) - \frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, X_j \rangle. \end{aligned} \quad (5.20)$$

To bound the empirical processes in the right hand side, we again use the bounds of steps 1–3 in the proof of Theorem 5. The bound of Step 1 yields

$$\left| \frac{1}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right) \right| \leq 16 \left( \sigma_{X \otimes X} \sqrt{\frac{t + \log(2m^2)}{n}} \bigvee U^2 \frac{t + \log(2m^2)}{n} \right)$$

and it follows from the bound of Step 2 that

$$\left| \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) \right| \leq 8U \|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{t + \log(2m)}{n}} \sqrt{16U^2 \frac{t + \log(2m)}{n}}.$$

Instead of more complicated derivation of Step 3, we now use noncommutative and classical Bernstein's inequalities to get that with probability at least  $1 - 2e^{-t}$

$$\begin{aligned} \left| n^{-1} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, X_j \rangle \right| &\leq \|\hat{\rho}^\varepsilon - S\|_1 \left\| n^{-1} \sum_{j=1}^n \xi_j X_j \right\| \leq 2 \left\| n^{-1} \sum_{j=1}^n \xi_j (X_j - \mathbb{E}X) \right\| + \\ 2\|\mathbb{E}X\| \left\| n^{-1} \sum_{j=1}^n \xi_j \right\| &\leq 4(\sigma_\xi \sigma_X + \|\mathbb{E}X\|) \sqrt{\frac{t + \log(2m)}{n}} \sqrt{12c_\xi U \frac{t + \log(2m)}{n}}. \end{aligned}$$

Using these inequalities, we derive from (5.20) that with some numerical constant  $C > 0$  and with probability at least  $1 - 4e^{-t}$ ,

$$\begin{aligned} \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; S) &\leq \|S - \rho\|_{L_2(\Pi)}^2 + \varepsilon \|\log S\| + \\ + C \left[ U \|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{t_m}{n}} + (\sigma_{X \otimes X} \vee \sigma_\xi \sigma_X \vee \|\mathbb{E}X\|) \sqrt{\frac{t_m}{n}} \vee (c_\xi U \vee U^2) \frac{t_m}{n} \right], \end{aligned} \quad (5.21)$$

which implies the result in the case when  $\|\log S\| \leq \log \Gamma$ . To finish the proof, it is enough, given an arbitrary  $S \in \mathcal{S}$  (even such that  $\log S$  does not exist), to apply bound (5.21) to  $S_\delta = (1 - \delta)S + \delta \frac{I_m}{m}$ , where  $\delta \in (0, 1)$ . Clearly,  $\|\log S_\delta\| \leq \log \frac{m}{\delta}$  and we also have  $\|S - S_\delta\|_{L_2(\Pi)}^2 \leq 4\delta^2 \mathbb{E}\|X\|^2$  (see the proof of Proposition 5). Taking  $\delta := \frac{\sqrt{\varepsilon}}{\mathbb{E}^{1/2}\|X\|^2} \wedge 1$ , it is easy to complete the proof in the case when  $\|\log S\| \geq \log \Gamma$ .

□

**Proof of Theorem 6.** Note that similarly to  $\rho^\varepsilon$ ,  $\hat{\rho}^\varepsilon$  is also a matrix of full rank and  $\log \hat{\rho}^\varepsilon$  is well defined. By necessary conditions of extrema in convex problems (1.2) and (4.1), we have  $DL_n(\hat{\rho}^\varepsilon; \hat{\rho}^\varepsilon - \rho^\varepsilon) \leq 0$  and  $DL(\rho^\varepsilon; \hat{\rho}^\varepsilon - \rho^\varepsilon) \geq 0$ . Subtracting the second inequality from the first one yields

$$DL(\hat{\rho}^\varepsilon; \hat{\rho}^\varepsilon - \rho^\varepsilon) - DL(\rho^\varepsilon; \hat{\rho}^\varepsilon - \rho^\varepsilon) \leq DL(\hat{\rho}^\varepsilon; \hat{\rho}^\varepsilon - \rho^\varepsilon) - DL_n(\hat{\rho}^\varepsilon; \hat{\rho}^\varepsilon - \rho^\varepsilon). \quad (5.22)$$

By a simple algebra already used in the proof of Theorem 5, this easily leads to the following bound:

$$2\|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; \rho^\varepsilon) \leq 2\mathbb{E} \langle \hat{\rho}^\varepsilon - \rho, X \rangle \langle \hat{\rho}^\varepsilon - \rho^\varepsilon, X \rangle - 2n^{-1} \sum_{j=1}^n (\langle \hat{\rho}^\varepsilon, X_j \rangle - Y_j) \langle \hat{\rho}^\varepsilon - \rho^\varepsilon, X_j \rangle,$$

which can be further rewritten as

$$2\|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; \rho^\varepsilon) \leq -\frac{2}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - \rho^\varepsilon, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - \rho^\varepsilon, X \rangle^2 \right) - \quad (5.23)$$

$$\frac{2}{n} \sum_{j=1}^n \left( \langle \rho^\varepsilon - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - \rho^\varepsilon, X_j \rangle - \mathbb{E} \langle \rho^\varepsilon - \rho, X \rangle \langle \hat{\rho}^\varepsilon - \rho^\varepsilon, X \rangle \right) - \frac{2}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - \rho^\varepsilon, X_j \rangle.$$

We use the bounds of steps 1–3 of the proof of Theorem 5 with  $S = \rho^\varepsilon$  to control each term in the right hand side of (5.23). Substituting these bounds in (5.23), we get the following inequality that holds with probability at least  $1 - 5e^{-t}$  :

$$2\|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; \rho^\varepsilon) \leq \quad (5.24)$$

$$8 \left( \sigma_{X \otimes X} \sqrt{\frac{t + \log(2m^2)}{n}} \bigvee U^2 \frac{t + \log(2m^2)}{n} \right) \|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_1^2 +$$

$$16U \|\rho^\varepsilon - \rho\|_{L_2(\Pi)} \sqrt{\frac{t + \log(2m)}{n}} \bigvee 16U^2 \|\rho^\varepsilon - \rho\|_1 \frac{t + \log(2m)}{n} +$$

$$4 \|P_{L^\perp}(\hat{\rho}^\varepsilon - \rho^\varepsilon) P_{L^\perp}\|_1 \left[ \sigma_\xi(\sigma_X + \|\mathbb{E}X\|) \sqrt{\frac{t + \log(2m)}{n}} \bigvee 2c_\xi U \frac{t + \log(2m)}{n} \right] +$$

$$C \left[ \sigma_\xi \beta(L) \|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)} \sqrt{\frac{mr}{n}} + \sigma_\xi \beta(L) \|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)} \sqrt{\frac{\tau_n}{n}} + c_\xi U \frac{\tau_n}{n} \right].$$

Arguing exactly as in the proof of Theorem 5, we can simplify (5.24) to get

$$2\|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; \rho^\varepsilon) \leq \quad (5.25)$$

$$16U \|\rho^\varepsilon - \rho\|_{L_2(\Pi)} \sqrt{\frac{t + \log(2m)}{n}} \bigvee 16U^2 \|\rho^\varepsilon - \rho\|_1 \frac{t + \log(2m)}{n} +$$

$$12 \|P_{L^\perp} \rho^\varepsilon P_{L^\perp}\|_1 \left[ \sigma_\xi(\sigma_X + \|\mathbb{E}X\|) \sqrt{\frac{t + \log(2m)}{n}} \bigvee 2c_\xi U \frac{t + \log(2m)}{n} \right] +$$

$$C \left[ \sigma_\xi \beta(L) \|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)} \sqrt{\frac{mr}{n}} + \sigma_\xi \beta(L) \|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)} \sqrt{\frac{\tau_n}{n}} + c_\xi U \frac{\tau_n}{n} \right].$$

It is easy now to solve this for  $\|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)}$  and to derive the following explicit bound on the random error that holds with probability at least  $1 - 5e^{-t}$  and with some numerical constant  $C > 0$  :

$$\|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; \rho^\varepsilon) \leq C \left[ \sigma_\xi^2 \beta^2(L) \frac{mr + \tau_n}{n} \bigvee c_\xi U \frac{\tau_n}{n} \bigvee \quad (5.26)$$

$$U \|\rho^\varepsilon - \rho\|_{L_2(\Pi)} \sqrt{\frac{t + \log(2m)}{n}} \bigvee U^2 \|\rho^\varepsilon - \rho\|_1 \frac{t + \log(2m)}{n} \bigvee$$

$$\|P_{L^\perp} \rho^\varepsilon P_{L^\perp}\|_1 \left( \sigma_\xi(\sigma_X \vee \|\mathbb{E}X\|) \sqrt{\frac{t + \log(2m)}{n}} \bigvee c_\xi U \frac{t + \log(2m)}{n} \right) \right],$$

which easily implies the result.  $\square$

**Example 1. Matrix completion (continuation).** Recall that, in this example,  $\{e_i : i = 1, \dots, m\}$  is the canonical basis of  $\mathbb{C}^m$  and the following set of Hermitian matrices forms an orthonormal basis of  $\mathbb{M}_m(\mathbb{C})$  (the matrix completion basis):

$$\left\{ e_i \otimes e_i : i = 1, \dots, m \right\} \cup \left\{ \frac{1}{\sqrt{2}}(e_i \otimes e_j + e_j \otimes e_i) : 1 \leq i < j \leq m \right\} \\ \cup \left\{ \frac{i}{\sqrt{2}}(e_i \otimes e_j - e_j \otimes e_i) : 1 \leq i < j \leq m \right\}.$$

Assume that  $X$  is sampled at random from this basis. Recall that in this case, for all matrices  $A$ ,  $\|A\|_{L_2(\Pi)}^2 = m^{-2}\|A\|_2^2$ . Obviously,  $\|e_i \otimes e_i\| = 1$ ,  $i = 1, \dots, m$  and, for all  $i < j$ ,

$$\left\| \frac{1}{\sqrt{2}}(e_i \otimes e_j + e_j \otimes e_i) \right\| = \frac{1}{\sqrt{2}}, \quad \left\| \frac{i}{\sqrt{2}}(e_i \otimes e_j - e_j \otimes e_i) \right\| = \frac{1}{\sqrt{2}}.$$

Therefore,  $\|X\| \leq U = 1$ . We also have

$$\sigma_X^2 \leq \|\mathbb{E}X^2\| = \sup_{v \in \mathbb{C}^m, |v|=1} \mathbb{E}\langle X^2 v, v \rangle = \sup_{v \in \mathbb{C}^m, |v|=1} \mathbb{E}\langle X v, X v \rangle = \sup_{v \in \mathbb{C}^m, |v|=1} \mathbb{E}|X v|^2.$$

Note that, if  $X = e_i \otimes e_i$ ,  $i = 1, \dots, m$ , then  $|X v|^2 = |e_i \langle e_i, v \rangle|^2 = |\langle e_i, v \rangle|^2$ . If  $X = \frac{1}{\sqrt{2}}(e_i \otimes e_j + e_j \otimes e_i)$ ,  $i < j$ , then

$$|X v|^2 = \frac{1}{2}|e_i \langle e_j, v \rangle + e_j \langle e_i, v \rangle|^2 = \frac{1}{2}(|\langle e_j, v \rangle|^2 + |\langle e_i, v \rangle|^2)$$

and, similarly, if  $X = \frac{i}{\sqrt{2}}(e_i \otimes e_j - e_j \otimes e_i)$ ,  $i < j$ , then also  $|X v|^2 = \frac{1}{2}(|\langle e_j, v \rangle|^2 + |\langle e_i, v \rangle|^2)$ . Therefore, for  $|v| = 1$ ,

$$\mathbb{E}|X v|^2 = m^{-2} \sum_{i=1}^m |\langle e_i, v \rangle|^2 + 2m^{-2} \frac{1}{2} \sum_{i < j} (|\langle e_j, v \rangle|^2 + |\langle e_i, v \rangle|^2) \leq \\ m^{-2}|v|^2 + m^{-2}m(|v|^2 + |v|^2) \leq 3m^{-1},$$

which implies that  $\sigma_X \leq \frac{\sqrt{3}}{\sqrt{m}}$ . By a similar simple computation,  $\sigma_{X \otimes X} \leq \frac{4}{\sqrt{m}}$ . Now we can derive the following corollary of Theorem 5. Let

$$\varepsilon_{n,m} := (\sigma_\xi m^{-1/2} \vee m^{-1/2}) \sqrt{\frac{t_m}{n}} \bigvee (c_\xi \vee 1) \frac{t_m}{n}$$

and let  $\varepsilon = D\varepsilon_{n,m}$  for a sufficiently large constant  $D > 0$ .

**Corollary 1** *There exists a numerical constant  $C > 0$  such that the following holds. For all  $t > 0$ , for all  $\lambda > 0$ , for all sufficiently large  $D$  and for  $\varepsilon = D\varepsilon_{n,m}$ , for all matrices  $S \in \mathcal{S}$  of rank  $r$ , with probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq (1 + \lambda)\|S - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ D^2 \left( (\sigma_\xi^2 \vee 1) \frac{rmt_m}{n} \vee \right. \right. \\ &\quad \left. \left. (c_\xi^2 \vee 1) \frac{rm^2 t_m^2}{n^2} \right) \log^2(mn) \vee \sigma_\xi^2 \frac{\tau_n}{n} \vee c_\xi \frac{\tau_n \vee t_m}{n} \vee \frac{t_m}{n} \right]. \end{aligned} \quad (5.27)$$

**Proof.** First observe that for all matrices  $S \in \mathcal{S}$  of full rank (for which  $\log S$  exists) and for all subspaces  $L \subset \mathbb{C}^m$  with  $\dim(L) = r$ , we have, with probability at least  $1 - e^{-t}$  and with an arbitrary  $\lambda > 0$

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq (1 + \lambda/2)\|S - \rho\|_{L_2(\Pi)}^2 + \frac{2C}{\lambda} \left[ a^2(\log S) \left( (\sigma_\xi^2 \vee 1) \frac{t_m}{mn} + (c_\xi^2 \vee 1) \frac{t_m^2}{n} \right) \vee \right. \\ &\quad \left. \sigma_\xi^2 \frac{mr + \tau_n}{n} \vee \sigma_\xi m^{-1/2} \|P_{L^\perp} S P_{L^\perp}\|_1 \sqrt{\frac{t_m}{n}} \vee c_\xi \frac{\tau_n \vee t_m}{n} \vee \frac{t_m}{n} \right]. \end{aligned} \quad (5.28)$$

This immediately follows from Theorem 5 since, in the case under consideration,  $\beta(L) = 1$ ,  $\sigma_X \leq 3^{1/2}m^{-1/2}$ ,  $\sigma_{X \otimes X} \leq 4m^{-1/2}$ ,  $U = 1$ . Note also that in this case  $\Lambda(L) = m$  (recall the definition of  $\Lambda(L)$  given before Proposition 5) and

$$a(\log S) \leq m \inf_c \|\log S + cI_m\|_2.$$

Suppose now that  $S \in \mathcal{S}$  is an arbitrary oracle of rank  $r$ . Then there exists a subspace  $L$  of dimension  $r$  such that  $P_{L^\perp} S P_{L^\perp} = 0$ . We will use bound (5.28) for  $S_\delta := (1 - \delta)S + \delta \frac{I_m}{m}$ , where  $\delta = \varepsilon \wedge 1$ , as we did in the proof of Proposition 5. As in this proof, we have, for some constant  $C_1 > 0$ ,

$$a(\log S_\delta) \leq m\sqrt{r} \log \left( 1 + \frac{m}{\delta} \right) \leq C_1 m\sqrt{r} \log(mn)$$

and

$$\|S - S_\delta\|_{L_2(\Pi)}^2 \leq 4\delta^2 \mathbb{E}\|X\|^2 \leq 4\delta^2 \leq 4\varepsilon^2.$$

Finally, note that

$$\|P_{L^\perp} S_\delta P_{L^\perp}\|_1 \leq (1 - \delta)\|P_{L^\perp} S P_{L^\perp}\|_1 + \delta\|P_{L^\perp} (I_m/m) P_{L^\perp}\|_1 \leq \delta \leq \varepsilon.$$

Substituting these bounds in (5.28) (with  $S$  replaced by  $S_\delta$ ) and bounding  $\|S_\delta - \rho\|_{L_2(\Pi)}^2$  in terms of  $\|S - \rho\|_{L_2(\Pi)}^2$  and  $\|S_\delta - S\|_{L_2(\Pi)}^2$  (similarly to what was done in the proof of

Proposition 5), it is easy to derive (5.27) from (5.28). Note that we can drop the term  $\sigma_\xi^2 \frac{mr}{n}$  since it is dominated by  $(\sigma_\xi^2 \vee 1) \frac{rmt_m}{n} \log^2(mn)$ .

□

Similarly, it is easy to obtain another corollary where the  $L_2(\Pi)$ -error of estimator  $\hat{\rho}^\varepsilon$  is controlled in terms of Gibbs oracles. Recall the notations at the end of Section 4 and also denote  $\Gamma_r := \|H_{\leq r}\|_2^2 = \sum_{k=1}^r \gamma_k^2$ .

**Corollary 2** *There exists a numerical constant  $C > 0$  such that the following holds. For all  $t > 0$ , for all  $\lambda > 0$ , for all sufficiently large  $D$  and for  $\varepsilon = D\varepsilon_{n,m}$ , for all Hermitian matrices  $H$  and for all  $r \leq m$ , with probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq (1 + \lambda) \|\rho_H - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ \frac{\delta_r^2(H)}{m^2} \bigvee D^2 \left( (\sigma_\xi^2 \vee 1) \frac{\Gamma_r m t_m}{n} \bigvee \right. \right. \\ &\quad \left. \left. (c_\xi^2 \vee 1) \frac{\Gamma_r m^2 t_m^2}{n^2} \right) \bigvee \sigma_\xi^2 \frac{mr + \tau_n}{n} \bigvee c_\xi \frac{\tau_n \vee t_m}{n} \bigvee \frac{t_m}{n} \right]. \end{aligned} \quad (5.29)$$

**Example 2. Pauli basis (continuation).** We now turn to another example described in the Introduction, the example of the Pauli basis. Recall that in this case  $m = 2^k$  and we are considering the basis of the space  $\mathbb{M}_{2^k}(\mathbb{C})$  that consists of all matrices of the form  $W_{i_1} \otimes \cdots \otimes W_{i_k}$ ,  $W_i = \frac{1}{\sqrt{2}} \sigma_i$ ,  $i = 1, \dots, 4$  being normalized  $2 \times 2$  Pauli matrices. Note that  $\|W_i\|_2 = 1$  and  $\|W_i\| = \frac{1}{\sqrt{2}}$ . The design variable  $X$  is picked at random from this basis. We still have  $\|A\|_{L_2(\Pi)}^2 = m^{-2} \|A\|_2^2$ . However, now

$$\|W_{i_1} \otimes \cdots \otimes W_{i_k}\| = \|W_{i_1}\| \cdots \|W_{i_k}\| = \left( \frac{1}{\sqrt{2}} \right)^k = 2^{-k/2} = m^{-1/2}$$

implying that  $\|X\| = m^{-1/2}$  and  $U = m^{-1/2}$ . To state a corollary of Theorem 5 in this case, we take  $\varepsilon := D\varepsilon_{n,m}$ , where

$$\varepsilon_{n,m} := (\sigma_\xi m^{-1/2} \vee m^{-1}) \sqrt{\frac{t_m}{n}} \bigvee (c_\xi m^{-1/2} \vee m^{-1}) \frac{t_m}{n}.$$

The following results are similar to corollaries 1 and 2.

**Corollary 3** *There exists a numerical constant  $C > 0$  such that the following holds. For all  $t > 0$ , for all  $\lambda > 0$ , for all sufficiently large  $D > 0$  and for  $\varepsilon = D\varepsilon_{n,m}$ , for all matrices  $S \in \mathcal{S}$  of rank  $r$ , with probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq (1 + \lambda) \|S - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ D^2 \left( (\sigma_\xi^2 \vee m^{-1}) \frac{rmt_m}{n} \bigvee \right. \right. \\ &\quad \left. \left. (c_\xi^2 \vee m^{-1}) \frac{rmt_m^2}{n^2} \right) \log^2(mn) \bigvee \sigma_\xi^2 \frac{\tau_n}{n} \bigvee c_\xi m^{-1/2} \frac{\tau_n \vee t_m}{n} \bigvee \frac{t_m}{mn} \right]. \end{aligned} \quad (5.30)$$



**Corollary 4** *There exists a numerical constant  $C > 0$  such that the following holds. For all  $t > 0$ , for all  $\lambda > 0$ , for all sufficiently large  $D$  and for  $\varepsilon = D\varepsilon_{n,m}$ , for all Hermitian matrices  $H$  and for all  $r \leq m$ , with probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq (1 + \lambda)\|\rho_H - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ \frac{\delta_r^2(H)}{m^2} \vee D^2 \left( (\sigma_\xi^2 \vee m^{-1}) \frac{\Gamma_r m t_m}{n} \vee \right. \right. \\ &\quad \left. \left. (c_\xi^2 \vee m^{-1}) \frac{\Gamma_r m^2 t_m^2}{n^2} \right) \vee \sigma_\xi^2 \frac{mr + \tau_n}{n} \vee c_\xi m^{-1/2} \frac{\tau_n \vee t_m}{n} \vee \frac{t_m}{mn} \right]. \end{aligned} \quad (5.31)$$

Note that the bounds of corollaries 1-4 can be also proved in the case when the noise is unbounded, in particular, Gaussian (see the remark after Theorem 6). For the Pauli basis, this immediately leads to Theorem 3 stated in the Introduction.

## 6 Oracle Inequalities: Subgaussian Design Case

In this section, we turn to the case of *subgaussian design matrices*. More precisely, we assume that  $X$  is a Hermitian random matrix with distribution  $\Pi$  such that, for some constant  $b_0 > 0$  and for all Hermitian matrices  $A \in \mathbb{M}_m(\mathbb{C})$ ,  $\langle A, X \rangle$  is a subgaussian random variable with parameter  $b_0 \|A\|_{L_2(\Pi)}$ . This implies that  $\mathbb{E}X = 0$  and, for some constant  $b_1 > 0$ ,

$$\left\| \langle A, X \rangle \right\|_{\psi_2} \leq b_1 \|A\|_{L_2(\Pi)}, \quad A \in \mathbb{M}_m(\mathbb{C}). \quad (6.1)$$

In addition to this, assume that, for some constant  $b_2 > 0$ ,

$$\|A\|_{L_2(\Pi)} = \left\| \langle A, X \rangle \right\|_{L_2(\Pi)} \leq b_2 \|A\|_2, \quad A \in \mathbb{M}_m(\mathbb{C}). \quad (6.2)$$

A Hermitian random matrix  $X$  satisfying the above conditions will be called a *subgaussian* matrix. Moreover, if  $X$  also satisfies the condition

$$\|A\|_{L_2(\Pi)}^2 = \mathbb{E}|\langle A, X \rangle|^2 = \|A\|_2^2, \quad A \in \mathbb{M}_m(\mathbb{C}), \quad (6.3)$$

then it will be called an *isotropic subgaussian* matrix. As it was already mentioned in the introduction, the last class of matrices includes such examples as Gaussian and Rademacher design matrices. It easily follows from the basic properties of Orlicz norms (see, e.g., van der Vaart and Wellner (1996), p. 95) that for subgaussian matrices  $\|A\|_{L_p(\Pi)} = \mathbb{E}^{1/p} \left| \langle A, X \rangle \right|^p \leq c_p b_1 b_2 \|A\|_2^2$  and  $\|A\|_{\psi_1} := \left\| \langle A, X \rangle \right\|_{\psi_1} \leq c b_1 b_2 \|A\|_2$ ,  $A \in \mathbb{M}_m(\mathbb{C})$ ,  $p \geq 1$ , with some numerical constants  $c_p > 0$  and  $c > 0$ .

The following is a version of a well known fact (see, e.g., Rudelson and Vershynin (2010), Proposition 2.4).

**Proposition 7** *Let  $X$  be a subgaussian  $m \times m$  matrix. Then, there exists a constant  $B > 0$  such that*

$$\left\| \|X\| \right\|_{\psi_2} \leq B\sqrt{m}.$$

**Proof.** Let  $M \subset S^{m-1} := \{u \in \mathbb{C}^m : |u| = 1\}$  be an  $\varepsilon$ -net of the unit sphere in  $\mathbb{C}^m$  of the smallest cardinality. It is easy to see that  $\text{card}(M) \leq (1 + 2/\varepsilon)^m$  and

$$\|X\| = \sup_{u,v \in S^{m-1}} \langle Xu, v \rangle \leq (1 - \varepsilon)^{-2} \max_{u,v \in M} \langle Xu, v \rangle.$$

Take  $\varepsilon = 1/2$ . Using standard bounds for Orlicz norms of a maximum (see, e.g., van der Vaart and Wellner (1996), Lemma 2.2.2), we get that, with some constants  $C_1, C_2, B > 0$ ,

$$\begin{aligned} \left\| \|X\| \right\|_{\psi_2} &\leq 4 \left\| \max_{u,v \in M} \langle Xu, v \rangle \right\|_{\psi_2} \leq C_1 \psi_2^{-1}(\text{card}^2(M)) \max_{u,v \in M} \left\| \langle Xu, v \rangle \right\|_{\psi_2} \leq \\ &C_2 \sqrt{\log \text{card}(M)} \max_{u,v \in M} \left\| \langle X, u \otimes v \rangle \right\|_{\psi_2} \leq C_2 \sqrt{\log \text{card}(M)} \max_{u,v \in M} \|u \otimes v\|_2 \leq B\sqrt{m}. \end{aligned}$$

□

Below, we give oracle inequalities and random error bounds in the subgaussian design case. We will use the following notations. Given  $t > 0$ , let

$$t_m := t + \log(2m), \quad \tau_n := t + \log \log_2(2n), \quad \text{and} \quad t_{n,m} := \tau_n \log n \vee t_m.$$

Also, denote  $c_\xi := \|\xi\|_{\psi_2} \log \frac{\|\xi\|_{\psi_2}}{\sigma_\xi}$  and let

$$\varepsilon_{n,m} := \sigma_\xi \sqrt{\frac{mt_m}{n}} \vee c_\xi \frac{\sqrt{mt_m}}{n}$$

(clearly, we assume here that the noise has a bounded  $\psi_2$ -norm).

**Theorem 7** *There exist constants  $C > 0, c > 0$  such that the following holds. For all  $t > 0$  and  $\lambda > 0$  such that  $\tau_n \leq c\lambda^2 n$ , for all  $S \in \mathcal{S}$  and for all  $\varepsilon \in [0, 1]$ , with probability at least  $1 - e^{-t}$*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 &\leq (1 + \lambda) \|S - \rho\|_{L_2(\Pi)}^2 + C \left[ \varepsilon \left( \|\log S\| \wedge \log \frac{m}{\varepsilon} \right) \vee \sigma_\xi \sqrt{\frac{mt_m}{n}} \vee \right. \\ &\left. \frac{mt_m}{n\lambda} \vee (c_\xi \vee \sqrt{m}) \frac{\sqrt{mt_{n,m}}}{n} \right] \end{aligned} \quad (6.4)$$

and

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq (1 + \lambda) \|S - \rho\|_{L_2(\Pi)}^2 + C \left[ \varepsilon \left( \|\log S\| \wedge \log \frac{m}{\varepsilon} \right) \vee \sigma_\xi \sqrt{\frac{mt_m}{n}} \vee \right. \\ &\left. \frac{mt_m}{n\lambda} \vee (c_\xi \vee \sqrt{m}) \frac{\sqrt{mt_{n,m}}}{n} \right]. \end{aligned} \quad (6.5)$$

In particular,

$$\|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 \leq C \left[ \varepsilon \left( \|\log \rho\| \wedge \log \frac{m}{\varepsilon} \right) \vee \sigma_\xi \sqrt{\frac{mt_m}{n}} \vee (c_\xi \vee \sqrt{m}) \frac{\sqrt{mt_{n,m}}}{n} \right].$$

We now turn to more subtle oracle inequalities that take into account low rank properties of oracles  $S \in \mathcal{S}$ .

**Theorem 8** *There exist numerical constants  $C > 0, D > 0, c > 0$  such that the following holds. For all  $t > 0$  and  $\lambda > 0$  such that  $\tau_n \leq c\lambda^2 n$ , for all  $\varepsilon \geq D\varepsilon_{n,m}$ , for all subspaces  $L \subset \mathbb{C}^m$  with  $\dim(L) := r$  and for all  $S \in \mathcal{S}$ , with probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 + \frac{\varepsilon}{4} K(\hat{\rho}^\varepsilon; S) &\leq (1 + \lambda) \|S - \rho\|_{L_2(\Pi)}^2 + \\ \frac{C}{\lambda} \left[ a^2 (\log S) \varepsilon^2 \vee \sigma_\xi^2 \beta^2(L) \frac{mr + \tau_n}{n} \vee \sigma_\xi \|P_{L^\perp} S P_{L^\perp}\|_1 \sqrt{\frac{mt_m}{n}} \vee (c_\xi \vee \sqrt{m}) \frac{\sqrt{mt_{n,m}}}{n} \right] \end{aligned} \quad (6.6)$$

and

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq (1 + \lambda) \|S - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ a^2 (\log S) \varepsilon^2 \vee \sigma_\xi^2 \frac{mr + \tau_n}{n} \vee \right. \\ &\left. \sigma_\xi \|P_{L^\perp} S P_{L^\perp}\|_1 \sqrt{\frac{mt_m}{n}} \vee (c_\xi \vee \sqrt{m}) \frac{\sqrt{mt_{n,m}}}{n} \right]. \end{aligned} \quad (6.7)$$

Similarly to the previous section, we also derived bounds on the random error  $\|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)}^2$ .

**Theorem 9** *There exist numerical constants  $C > 0, D > 0, c > 0$  such that the following holds. Under the assumption that  $\tau_n \leq cn$ , for all  $t > 0$ , for all  $\varepsilon \geq D\varepsilon_{n,m}$  and for all subspaces  $L \subset \mathbb{C}^m$  with  $\dim(L) := r$ , with probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho^\varepsilon\|_{L_2(\Pi)}^2 + \varepsilon K(\hat{\rho}^\varepsilon; \rho^\varepsilon) &\leq C \left[ \sigma_\xi^2 \beta^2(L) \frac{mr + \tau_n}{n} \vee \sigma_\xi \|P_{L^\perp} \rho^\varepsilon P_{L^\perp}\|_1 \sqrt{\frac{mt_m}{n}} \vee \right. \\ &\left. \|\rho^\varepsilon - \rho\|_{L_2(\Pi)} \sqrt{\frac{mt_m}{n}} \vee (c_\xi \vee \sqrt{m}) \frac{\sqrt{mt_{n,m}}}{n} \right]. \end{aligned} \quad (6.8)$$

We will give only the proof of Theorem 8.

**Proof.** It follows the lines of the proof of Theorem 5 very closely. The main changes are in the bounds of steps 1–3 of this proof that have to be modified in the subgaussian design case. The rest of the proof is straightforward.

In Step 1, we have to bound the following quantity:

$$\frac{1}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right).$$

To this end, we will study the empirical process

$$\Delta_n(\delta) := \sup_{f \in \mathcal{F}_\delta} \left| n^{-1} \sum_{j=1}^n (f^2(X_j) - P f^2) \right|,$$

where  $\mathcal{F}_\delta := \{ \langle S_1 - S_2, \cdot \rangle : S_1, S_2 \in \mathcal{S}, \|S_1 - S_2\|_{L_2(\Pi)} \leq \delta \}$ . Clearly,

$$\left| \frac{1}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right) \right| \leq \Delta_n(\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}).$$

Our goal is to obtain an upper bound on  $\Delta_n(\delta)$  uniformly in  $\delta \in [(m/n)^{1/2}, 2b_2]$ . First we use a version of Talagrand's concentration inequality for empirical processes indexed by unbounded functions due to Adamczak (see subsection 3.2). It implies that with some constant  $C > 0$  and with probability at least  $1 - e^{-t}$

$$\Delta_n(\delta) \leq 2\mathbb{E}\Delta_n(\delta) + C\delta^2 \sqrt{\frac{t}{n}} + C \frac{mt \log n}{n}. \quad (6.9)$$

Here we used the following bounds on the uniform variance and on the envelope of the function class  $\mathcal{F}_\delta^2$ : for the uniform variance, with some constant  $c > 0$ ,

$$\begin{aligned} \sup_{f \in \mathcal{F}_\delta} (P f^4)^{1/2} &= \sup_{S_1, S_2 \in \mathcal{S}, \|S_1 - S_2\|_{L_2(\Pi)} \leq \delta} \mathbb{E}^{1/2} \langle S_1 - S_2, X \rangle^4 = \\ &= \sup_{S_1, S_2 \in \mathcal{S}, \|S_1 - S_2\|_{L_2(\Pi)} \leq \delta} \|S_1 - S_2\|_{L_4(\Pi)}^2 \leq c\delta^2, \end{aligned}$$

by the equivalence properties of the norms in Orlicz spaces. For the envelope,

$$\sup_{f \in \mathcal{F}_\delta} f^2(X) = \sup_{S_1, S_2 \in \mathcal{S}, \|S_1 - S_2\|_{L_2(\Pi)} \leq \delta} \langle S_1 - S_2, X \rangle^2 \leq 4\|X\|^2$$

and

$$\left\| \max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}_\delta} f^2(X_i) \right\|_{\psi_1} \leq c_1 \left\| \|X\|^2 \right\|_{\psi_1} \log n \leq c_2 \left\| \|X\| \right\|_{\psi_2}^2 \log n \leq c_3 m \log n,$$

for some constants  $c_1, c_2, c_3 > 0$ , where we used well known inequalities for maxima of random variables in Orlicz spaces (see, e.g., Lemma 2.2.2 in van der Vaart and Wellner (1996)).

To bound the expectation  $\mathbb{E}\Delta_n(\delta)$  we use a recent result by Mendelson (2010) (see subsection 3.2; in fact, even earlier result by Klartag and Mendelson (2005) with the  $\psi_2$ -diameter instead of  $\psi_1$ -diameter would suffice for our purposes). It gives

$$\mathbb{E}\Delta_n(\delta) \leq c \left[ \sup_{f \in \mathcal{F}_\delta} \|f\|_{\psi_1} \frac{\gamma_2(\mathcal{F}_\delta; \psi_2)}{\sqrt{n}} \vee \frac{\gamma_2^2(\mathcal{F}_\delta; \psi_2)}{n} \right] \quad (6.10)$$

with some constant  $c > 0$ . It follows from (6.1) that the  $\psi_1$  and  $\psi_2$ -norms of functions from the class  $\mathcal{F}_\delta$  can be bounded from above by a constant times the  $L_2(P)$ -norm. As a result,

$$\sup_{f \in \mathcal{F}_\delta} \|f\|_{\psi_1} \leq c\delta \quad (6.11)$$

and the following bound holds for Talagrand's generic chaining complexities:

$$\gamma_2(\mathcal{F}_\delta; \psi_2) \leq \gamma_2(\mathcal{F}_\delta; c\|\cdot\|_{L_2(\Pi)}), \quad (6.12)$$

where  $c$  is a constant. Let  $G$  be a symmetric real valued random matrix with independent centered Gaussian entries  $\{g_{ij}\}$  on the diagonal and above, where  $\mathbb{E}g_{ii}^2 = 1$  and  $\mathbb{E}g_{ij}^2 = \frac{1}{2}, i \neq j$ . Then, using condition (6.2), we have that, for some constant  $c_1 > 0$ ,

$$\mathbb{E}|\langle S_1, G \rangle - \langle S_2, G \rangle|^2 = \|S_1 - S_2\|_2^2 \geq c_1 \|S_1 - S_2\|_{L_2(\Pi)}^2,$$

and it easily follows from Talagrand's generic chaining bound that, for some constant  $C > 0$ ,

$$\gamma_2(\mathcal{F}_\delta; c\|\cdot\|_{L_2(\Pi)}) \leq C \mathbb{E} \sup_{S_1, S_2 \in \mathcal{S}, \|S_1 - S_2\|_{L_2(\Pi)} \leq \delta} |\langle S_1 - S_2, G \rangle| =: C\omega(G; \delta). \quad (6.13)$$

It follows from (6.10), (6.11), (6.12) and (6.13) that

$$\mathbb{E}\Delta_n(\delta) \leq C \left[ \delta \frac{\omega(G; \delta)}{\sqrt{n}} \vee \frac{\omega^2(G; \delta)}{n} \right]. \quad (6.14)$$

To bound  $\mathbb{E} \sup_{S_1, S_2 \in \mathcal{S}, \|S_1 - S_2\|_{L_2(\Pi)} \leq \delta} |\langle S_1 - S_2, G \rangle|$ , note that

$$\left| \langle S_1 - S_2, G \rangle \right| \leq \|S_1 - S_2\|_1 \|G\| \leq 2\|G\|,$$

and, by Proposition 7,

$$\omega(G; \delta) = \mathbb{E} \sup_{\rho_1, \rho_2 \in \mathcal{S}, \|\rho_1 - \rho_2\|_{L_2(\Pi)} \leq \delta} \left| \langle S_1 - S_2, G \rangle \right| \leq 2\mathbb{E}\|G\| \leq c\sqrt{m}.$$

Substituting this bound in (6.14) yields that, for some constant  $C > 0$ ,

$$\mathbb{E}\Delta_n(\delta) \leq C \left[ \delta \sqrt{\frac{m}{n}} \vee \frac{m}{n} \right] \quad (6.15)$$

and combining (6.15) with (6.9) gives that with probability at least  $1 - e^{-t}$

$$\Delta_n(\delta) \leq C \left[ \delta \sqrt{\frac{m}{n}} \vee \frac{m}{n} \vee \delta^2 \sqrt{\frac{t}{n}} \vee \frac{mt \log n}{n} \right]. \quad (6.16)$$

It is easy to make bound (6.16) uniform in  $\delta \in [(m/n)^{1/2}, 2b_2]$  by a simple discretization argument (as we did in Step 3 of the proof of Theorem 5). This leads to the following result: with probability at least  $1 - e^{-t}$ , for all  $\delta \in [(m/n)^{1/2}, 2b_2]$ ,

$$\Delta_n(\delta) \leq C \left[ \delta \sqrt{\frac{m}{n}} \vee \frac{m}{n} \vee \delta^2 \sqrt{\frac{\tau_n}{n}} \vee \frac{m\tau_n \log n}{n} \right], \quad (6.17)$$

where  $\tau_n = t + \log \log_2(2n)$ . Thus, with the same probability and with a proper choice of constant  $C > 0$

$$\begin{aligned} & \left| \frac{1}{n} \sum_{j=1}^n \left( \langle \hat{\rho}^\varepsilon - S, X_j \rangle^2 - \mathbb{E} \langle \hat{\rho}^\varepsilon - S, X \rangle^2 \right) \right| \leq \\ & C \left[ \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} \sqrt{\frac{m}{n}} \vee \frac{m}{n} \vee \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)}^2 \sqrt{\frac{\tau_n}{n}} \vee \frac{m\tau_n \log n}{n} \right] \end{aligned}$$

provided that  $\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} \in [(m/n)^{1/2}, 2b_2]$ .

Similarly to Step 2 of the proof of Theorem 5, we have to bound the expression

$$\begin{aligned} & \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) = \\ & \left\langle \hat{\rho}^\varepsilon - S, \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle X_j - \mathbb{E} \langle S - \rho, X \rangle X \right) \right\rangle. \end{aligned}$$

We use the bound

$$\begin{aligned} & \left| \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) \right| \leq \\ & \|\hat{\rho}^\varepsilon - S\|_1 \left\| \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle X_j - \mathbb{E} \langle S - \rho, X \rangle X \right) \right\| \end{aligned}$$

and Proposition 2 with  $\alpha = 1$ . Note that

$$\|\mathbb{E} \langle S - \rho, X \rangle^2 X^2\| \leq \mathbb{E} \langle S - \rho, X \rangle^2 \|X\|^2 \leq \mathbb{E}^{1/2} \langle S - \rho, X \rangle^4 \mathbb{E}^{1/2} \|X\|^4 \leq cm \|S - \rho\|_{L_2(\Pi)}^2$$

with a constant  $c > 0$ . Also,

$$\left\| \langle S - \rho, X \rangle X \right\|_{\psi_1} = \left\| \langle S - \rho, X \rangle \|X\| \right\|_{\psi_1} \leq c_1 \|\langle S - \rho, X \rangle\|_{\psi_2} \left\| \|X\| \right\|_{\psi_2} \leq c_2 \sqrt{m} \|S - \rho\|_{L_2(\Pi)}$$

with some constants  $c_1, c_2 > 0$ . Finally, note that

$$\|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} \leq b_2 \|S - \rho\|_2 \leq b_2 \|S - \rho\|_1 \|S - \rho\| \leq 4b_2,$$

since, for  $S, \rho \in \mathcal{S}$ ,  $\|S - \rho\|_1 \leq 2$  and  $\|S - \rho\| \leq 2$ . Using the fact  $\|\hat{\rho}^\varepsilon - S\|_1 \leq 2$ , Proposition 2 and the previous bounds imply that with probability at least  $1 - e^{-t}$  and with some constants  $C_1, C_2, C > 0$ ,

$$\begin{aligned} & \left| \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle \langle \hat{\rho}^\varepsilon - S, X_j \rangle - \mathbb{E} \langle S - \rho, X \rangle \langle \hat{\rho}^\varepsilon - S, X \rangle \right) \right| \leq \\ & \|\hat{\rho}^\varepsilon - S\|_1 \left\| \frac{1}{n} \sum_{j=1}^n \left( \langle S - \rho, X_j \rangle X_j - \mathbb{E} \langle S - \rho, X \rangle X \right) \right\| \leq \\ & C_1 \left[ \|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{m(t + \log(2m))}{n}} \vee \frac{\sqrt{m}(t + \log(2m))}{n} \|S - \rho\|_{L_2(\Pi)} \log \frac{C_1 \sqrt{m} \|S - \rho\|_{L_2(\Pi)}}{\sqrt{m} \|S - \rho\|_{L_2(\Pi)}} \right] \leq \\ & C \left[ \|S - \rho\|_{L_2(\Pi)} \sqrt{\frac{m(t + \log(2m))}{n}} \vee \frac{\sqrt{m}(t + \log(2m))}{n} \right]. \end{aligned}$$

We now modify the bounds of Step 3 of the proof of Theorem 5. We need to bound the following expression:

$$\frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, X_j \rangle = \left\langle P_{L^\perp} (\hat{\rho}^\varepsilon - S) P_{L^\perp}, \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\rangle + \frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, \mathcal{P}_L X_j \rangle.$$

As in the proof of Theorem 5,

$$\left| \left\langle P_{L^\perp} (\hat{\rho}^\varepsilon - S) P_{L^\perp}, \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\rangle \right| \leq \|P_{L^\perp} (\hat{\rho}^\varepsilon - S) P_{L^\perp}\|_1 \left\| \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\|.$$

By Proposition 2, it is easy to show that with probability at least  $1 - e^{-t}$ ,

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\| \leq \\ & C \left[ \sigma_\xi \|\mathbb{E} X^2\|^{1/2} \sqrt{\frac{t + \log(2m)}{n}} \vee \|\xi\|_{\psi_2} \left\| \|X\| \right\|_{\psi_2} \log \left( \frac{\|\xi\|_{\psi_2} \left\| \|X\| \right\|_{\psi_2}}{\sigma_\xi \sigma_X} \right) \frac{t + \log(2m)}{n} \right]. \end{aligned}$$

We replace  $\sigma_X, \|\mathbb{E}X^2\|^{1/2}$  and  $\|X\|_{\psi_2}$  by an upper bound  $c\sqrt{m}$  (see Proposition 7) which yields a simplified inequality

$$\left\| \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\| \leq C \left[ \sigma_\xi \sqrt{\frac{m(t + \log(2m))}{n}} \vee \|\xi\|_{\psi_2} \log\left(\frac{\|\xi\|_{\psi_2}}{\sigma_\xi}\right) \frac{\sqrt{m}(t + \log(2m))}{n} \right].$$

Hence, with probability at least  $1 - e^{-t}$ ,

$$\begin{aligned} & \left| \left\langle P_{L^\perp}(\hat{\rho}^\varepsilon - S)P_{L^\perp}, \frac{1}{n} \sum_{j=1}^n \xi_j P_{L^\perp} X_j P_{L^\perp} \right\rangle \right| \leq \\ & C \|P_{L^\perp}(\hat{\rho}^\varepsilon - S)P_{L^\perp}\|_1 \left[ \sigma_\xi \sqrt{\frac{m(t + \log(2m))}{n}} \vee \|\xi\|_{\psi_2} \log\left(\frac{\|\xi\|_{\psi_2}}{\sigma_\xi}\right) \frac{\sqrt{m}(t + \log(2m))}{n} \right]. \end{aligned}$$

The remaining term  $\frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, P_L X_j \rangle$  is bounded exactly as in Step 3 of the proof of Theorem 5 with the use of Adamczak's (2008) version of Talagrand's concentration inequality. This leads to the following bound: with probability at least  $1 - e^{-t}$ ,

$$\left| \frac{1}{n} \sum_{j=1}^n \xi_j \langle \hat{\rho}^\varepsilon - S, P_L X_j \rangle \right| \leq C \left[ \sigma_\xi \beta(L) \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} \sqrt{\frac{mr}{n}} + \sigma_\xi \|\hat{\rho}^\varepsilon - S\|_{L_2(\Pi)} \sqrt{\frac{\tau_n}{n}} + \|\xi\|_{\psi_2} \frac{\sqrt{m} \tau_n \log n}{n} \right],$$

where  $\tau_n = t + \log \log_2(2n)$ .

□

For simplicity, we state the next corollaries (similar to corollaries 1 and 2) only in the case of subgaussian isotropic design. Recall that in this case  $\|\cdot\|_{L_2(\Pi)} = \|\cdot\|_2$  and  $\beta(L) = 1$ .

**Corollary 5** *There exist numerical constants  $C > 0, c > 0$  such that the following holds. For all  $t > 0$  and  $\lambda > 0$  such that  $\tau_n \leq c\lambda^2 n$ , for all sufficiently large  $D > 0$  and for  $\varepsilon = D\varepsilon_{n,m}$ , for all matrices  $S \in \mathcal{S}$  of rank  $r$ , with probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 & \leq (1 + \lambda) \|S - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ D^2 \left( \sigma_\xi^2 \frac{rmt_m}{n} \vee c_\xi^2 \frac{rmt_m^2}{n^2} \right) \log^2(mn) \vee \right. \\ & \left. \sigma_\xi^2 \frac{\tau_n}{n} \vee (c_\xi \vee \sqrt{m}) \frac{\sqrt{m} \tau_{n,m}}{n} \right]. \end{aligned} \quad (6.18)$$

**Corollary 6** *There exists numerical constants  $C > 0, c > 0$  such that the following holds. For all  $t > 0$  and for all  $\lambda > 0$  such that  $\tau_n \leq c\lambda^2 n$ , for all sufficiently large  $D$*



and for  $\varepsilon = D\varepsilon_{n,m}$ , for all Hermitian matrices  $H$  and for all  $r \leq m$ , with probability at least  $1 - e^{-t}$ ,

$$\begin{aligned} \|\hat{\rho}^\varepsilon - \rho\|_{L_2(\Pi)}^2 &\leq (1 + \lambda)\|\rho_H - \rho\|_{L_2(\Pi)}^2 + \frac{C}{\lambda} \left[ \delta_r^2(H) \vee D^2 \left( \sigma_\xi^2 \frac{\Gamma_r m t_m}{n} \vee \right. \right. \\ &\left. \left. c_\xi^2 \frac{\Gamma_r m t_m^2}{n^2} \right) \vee \sigma_\xi^2 \frac{mr + \tau_n}{n} \vee (c_\xi \vee \sqrt{m}) \frac{\sqrt{m} t_{n,m}}{n} \right]. \end{aligned} \quad (6.19)$$

In a special case of Gaussian noise, the bounds of the above corollaries can be simplified since in this case  $c_\xi \leq c\sigma_\xi$  for some numerical constant  $c$ . In particular, Corollary 5 immediately implies the bound of Theorem 2 in the Introduction. Both bounds of Theorem 1 follow from theorems 7 and 8.

## References

- [1] Adamczak, R. (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13, 34, 1000–1034.
- [2] Artiles, L.M., Gill, R. and Guta, M.I. (2004) An invitation to quantum tomography. *J. Royal Statistical Society, Ser. B*, v. 67, 1, 109–134.
- [3] Ahlswede, R. and Winter, A. (2002) Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48, 3, pp. 569–679.
- [4] Bhatia, R. (1997) *Matrix Analysis*. Springer, New York.
- [5] Candes, E. and Recht, B. (2009) Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6), 717–772.
- [6] Candes, E. and Tao, T. (2009) The power of convex relaxation: Near-optimal matrix completion. Technical Report.
- [7] Candes, E. and Plan, Y. (2009) Tight Oracle Bounds for Low-Rank Matrix Recovery from a Minimal Number of Random Measurements. Preprint.
- [8] Gross, D., Lou, Yo-Kai, Flammia, S.T., Becker, S. and Assert, J. (2009) Quantum State Tomography via compressed sensing. Preprint.
- [9] Gross, D. (2009) Recovering Low-Rank Matrices From Few Coefficients in Any Basis. Preprint.
- [10] Klartag, B. and Mendelson, S. (2005) Empirical Processes and Random Projections. *Journal of Functional Analysis*, 225(1), 229–245.
- [11] Klauck, H., Nayak, A., Ta-Shma, A. and Zuckerman, D. (2007) Interactions in Quantum Communication. *IEEE Transactions on Information Theory*, 53, 6, 1970–1982.
- [12] Koltchinskii, V. (2009) Sparse recovery in convex hulls via entropy penalization. *Annals of Statistics*, 37(3), 1332–1359.
- [13] Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces*. Springer.

- [14] Mendelson, S. (2010) Empirical processes with a bounded  $\psi_1$  diameter. Preprint.
- [15] Nielsen, M.A. and Chang, I.L. (2000) Quantum Computation and Quantum Information, Cambridge University Press.
- [16] Recht, B. (2009) A Simpler Approach to Matrix Completion. Preprint.
- [17] Rudelson, M. and Vershynin, R. (2010) Non-asymptotic theory of random matrices: extreme singular values. *Proceedings of the International Congress of Mathematicians*, Hyderabad, India.
- [18] Rohde, A. and Tsybakov, A. (2009) Estimation of high-dimensional low rank matrices. Preprint.
- [19] Simon, B. (1979) Trace Ideals and their Applications. Cambridge University Press.
- [20] Talagrand, M. (2005) The Generic Chaining. Springer.
- [21] van der Vaart, A. and Wellner, J. (1996) Weak Convergence and Empirical Processes. With Applications to Statistics. Springer.